BAYESIAN GLS REGRESSION, LEVERAGE, AND INFLUENCE FOR REGIONALIZATION OF HYDROLOGIC STATISTICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Andrea Gruber Veilleux

August 2011

BAYESIAN GLS REGRESSION, LEVERAGE, AND INFLUENCE FOR REGIONALIZATION OF HYDROLOGIC STATISTICS

Andrea Gruber Veilleux, Ph. D.

Cornell University 2011

The research presented in this dissertation develops new statistical techniques for estimating regional relationships of hydrologic statistics. These techniques include extensions of the quasi-analytic Bayesian Generalizes Least Squares (B-GLS) framework presented in Reis *et al*. [2005] and further developed by Gruber *et al.* [2007] and Gruber and Stedinger [2008]. Recent extensions include a Pseudo R_δ^2 and pseudo analysis of variance table, plus a range of model performance, diagnostic and goodness-of-fit statistic. This dissertation develops a more stable Bayesian WLS/GLS procedure with the corresponding measures of precision and model performance. Special attention is given to model performance criteria, and the meaning of and insight provided by alternative measures of leverage and influence.

Examples address development of regional skewness coefficients to improve flood frequency analysis in the United States. Large cross-correlations between annual peak discharges, coupled with relatively small model error variances, present difficulties in regional GLS skewness analyses. The B-GLS framework seeks to exploit the cross-correlations among the sample skewness estimates to obtain the best possible estimates of the model parameters. However, if the cross-correlations are large, the GLS estimators can become relatively complicated as a result of the effort to find the most efficient estimator of the parameters. Unfortunately, it appears that the precision of the cross-correlation estimates between any two particular sites is not of sufficient precision to justify the seemingly incorrect weights (both

positive and negative) that the B-GLS analysis generates. Thus, an alternate regression procedure using both Weighted Least Squares (WLS) and GLS is developed so that the regional skewness analysis can provide both stable and defensible results. This alternate regression framework, is applied to two different data sets from different parts of the United States: the State of California and the Southeastern United States, to develop regional skewness estimators for flood frequency analysis.

In addition, special attention is devoted to comparing and developing leverage and influence diagnostics statistics for GLS and WLS/GLS analyses, which can be used to identify rogue observations and to effectively address lack-of-fit when estimating hydrologic statistics.

BIOGRAPHICAL SKETCH

Andrea Marie Gruber was born June 8, 1984 in Fountain Valley, California. She grew up in sunny Southern California with her parents, brother, and sister. Andrea graduated cum laude from Cornell University in May 2006 with a B.S. in Civil and Environmental Engineering. In August 2009, Andrea received M.S. degree at Cornell in the department of Civil and Environmental Engineering with a concentration in Environmental and Water Resources Systems. After which she remained at Cornell to pursue her Ph.D. in the same field. On October 18, 2008, she married fellow Cornell engineering graduate student Michael Veilleux and took the name Andrea Gruber Veilleux.

to my husband mike

and my family,

evan, jane, aaron and allison

thanks for distracting me with distractions

ACKNOWLEDGMENTS

I owe thanks to many people who have provided encouragement and assistance as I worked to complete this dissertation. First and foremost, I would like to thank my research advisor and special committee chair, Professor Jery R. Stedinger. His expert guidance and support along with his endless enthusiasm for his work is inspiring. His valuable wisdom has been instrumental in fostering an incredibly challenging and rewarding learning experience and has served as great preparation as I get ready to start my own career. I would also like to thank my minor advisors and special committee members; Professor Wilfried Brutsaert for his counseling and insight and Professor Huaizhu Gao for expanding my horizons beyond hydrology.

I sincerely acknowledge the support provided by a Water Resources Institute Internship Award #07HQAGOI61 by the U.S. Geological Survey, U.S. Department of the Interior. I am grateful to Dr. Timothy Cohn for the knowledgeable input he provided regarding the research presented in this thesis. I would also like to thank my friend and colleague Dr. Ken Eng for assisting me in my research and for continually encouraging me to think big.

Most importantly, I would like to thank my family for always being there for me. I wouldn't be who I am today without the love and support of my parents Evan and Jane, my brother Aaron, and my sister Allison. And most of all to my best friend and husband Mike, thank you for helping me through the stressful days with your gourmet meals.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

The research presented in this dissertation develops new statistical techniques for estimating regional relationships of hydrologic statistics. These techniques include extensions of the Bayesian Generalizes Least Squares (B-GLS) framework presented in Reis *et al*. [2005]. Recent extensions include a Pseudo R_{δ}^2 and pseudo analysis of variance table, plus a range of model performance, diagnostic and goodness-of-fit statistics. In some cases, the B-GLS analysis proved to be unstable. This dissertation develops a more stable Bayesian WLS-GLS procedure with the corresponding measures of precision and model performance, and that new methodology is demonstrated with hydrologic data sets from California and the Southeastern U.S. Special attention is given to model performance criteria, and the meaning of and insight provided by alternative measures of leverage and influence.

Examples generally address development of regional skewness coefficients to improve flood frequency analysis in the United States. Flood frequency guidelines for the United States, specified in *Bulletin 17B*, recommend fitting the log-Pearson Type III (LP3) distribution to the series of annual flood maxima, in which the third moment of the distribution, the skewness coefficient, is combined with a regional skewness coefficient to improve its precision. The research presented here extends the quasi-analytic Bayesian analysis of the Generalized Least Squares (GLS) regional hydrologic regression framework introduced by Reis *et al*. [2005], and furthered developed by Gruber *et al.* [2007], Gruber and Stedinger [2008], and Veilleux [2009] to estimate more accurately and precisely regional hydrologic relationships. In particular,

examples consider the skewness coefficients. Large cross-correlations between annual peak discharges, coupled with relatively small model error variances, present difficulties in regional GLS skewness analyses. The Bayesian GLS (B-GLS) framework seeks to exploit the cross-correlations among the sample skewness estimates to obtain the best possible estimates of the model parameters. However, if the cross-correlations are large, the GLS estimators can become relatively complicated as a result of the effort to find the most efficient estimator of the parameters. Unfortunately, it appears that the precision of the cross-correlation estimates between any two particular sites is not of sufficient precision to justify the seemingly incorrect weights (both positive and negative) that the B-GLS analysis generates. Thus, an alternate regression procedure using both Weighted Least Squares (WLS) and GLS is developed so that the regional skewness analysis can provide both stable and defensible results.

In addition, special attention is devoted to comparing and developing leverage and influence diagnostics statistics for GLS and WLS/GLS analyses, which can be used to identify rogue observations and to effectively address lack-of-fit when estimating skewness coefficients or other hydrologic statistics. This alternate regression framework, which uses both B-WLS and B-GLS, is illustrated with two different data sets from different parts of the United States: the State of California and the Southeastern United States, to develop regional skewness estimators for use in flood frequency analysis.

Because of the focus in this dissertation on regional estimation of the skewness coefficient to improve flood frequency procedures in the United States, Section 1.1 below provides background on United States flood flow frequency procedures, highlights the importance of the skewness coefficient in determining estimates of flood quantiles, and clarifies the need for an improved regional skewness estimator. Section 1.2 presents a justification for a

Bayesian Generalized Least Squares framework for regional hydrologic regression analyses. Finally, Section 1.3 offers a detailed outline of the other chapters in this dissertation. Those chapters address leverage and influence for use in regression, a regional skewness model for the State of California based on annual maximum floods, and an extended B-WLS/B-GLS framework for estimating regional hydrologic parameters with an example using a data set from the Southeastern U.S.

1.1 United States Flood Flow Frequency Procedures and Regional Skew

As described recently in Griffis and Stedinger [2007c], Stedinger and Griffis [2008] and Veilleux [2009], the United States Water Resource Council published a series of guidelines entitled "A Uniform Technique for Determining Flood Flow Frequencies" in response to an increased social interest in managing flood loss and decreasing flood risk within the United States. The first version of these guidelines was published in 1967 [IACWD, 1982]. That document was followed by *Bulletin 17* in 1975, *Bulletin 17A* in 1976 and the current version, *Bulletin 17B*, published in 1982 [IACWD, 1982]. An update of the U.S. recommended flood frequency guidelines has not occurred since 1982, thus almost 35 years have passed without revisions. As argued by Stedinger and Griffis [2008], it is essential that the prescribed techniques in *Bulletin 17B* be updated to make use of recent advances in the field of flood frequency analysis, and to maintain the credibility of US Government procedures in this important and contentious area.

Veilleux [2009] observes that *Bulletin 17B* recommended use of the log-Pearson Type III (LP3) distribution to fit a series of annual maximum flood peaks to obtain a flood-frequency relationship. This distribution, in the specific case of flood frequency analysis, is described by

three moments: the mean, the standard deviation, and the skewness coefficient of the logarithms of the flow. The third moment, the skewness coefficient, is a measure of the asymmetry of the distribution or, in other words, the relative thickness of the tails of the distribution. The traditional sample estimator of the skewness coefficient is very sensitive to extreme events, such as large floods or unusually small values, as they cause a sample to be highly skewed, or asymmetrical. Thus, in flood frequency analysis, the skewness coefficient becomes significant because interest is focused on the right-hand tail of the distribution. However, the span of available years of recorded flood data at a given gauge site is usually too short (less than 120yrs, and often less than 35 years) to provide a highly reliable estimate of the skewness coefficient.

In order to improve the precision of the skewness estimator, *Bulletin 17B* advises combining a regional skew with the at-site skew estimator. A number of papers have described the possible use of a regional skew as well as its estimation and likely precision [Beard 1974; Griffis and Stedinger, 2007a,c; Hardison, 1975; IACWD, 1982; McCuen, 1979, 2001; Tasker, 1978]. Griffis and Stedinger [2009b, Appendix] show that the *Bulletin 17B* mean squared error (MSE) weighted skewness estimator results in the estimator with the smallest MSE provided that the regional skew is unbiased and independent of the at-site skew estimator. Griffis and Stedinger [2007a,c; 2009a] illustrate the value of a good regional skewness estimator in terms of the precision of flood quantile estimates.

When putting *Bulletin 17B* into practice, regional skew values may be obtained from the skew map included with the original *Bulletin*. This skew map is still used today, over 35 years later. The first edition of *Bulletin 17* states: "It is expected that Plate I [the skew map] will be revised as more data become available and more extensive studies are completed." (See text printed on back of map.) However, in spite of the tremendous advances in computing power

over the past few decades which support the Bayesian GLS regional hydrologic regression framework, the skew map has not been updated nationally, nor has a revision of *Bulletin 17B* been generated, though efforts are currently underway to do so [Stedinger and Griffis, 2008].

1.2 Regional Hydrologic Regression Analysis

Hydrologic studies at ungauged sites pose a challenge precisely because no flow records are available at those locations. Thus, there is a desire to develop regional hydrologic relationships based upon records available across a region, as noted by Veilleux [2009]. Also, at gauged sites, records can be too short to provide highly accurate at-site estimates of flood quantiles, low flows, and other regional hydrologic statistics. Thus, regional information can also be of use to improve the accuracy of estimates in these cases [IWACD, 1982, Section V.B.4]. One approach for relating data from gauged sites to ungauged sites is to derive empirical relationships between the hydrologic variable of interest and various measurable basin characteristics at the gauged sites using regional regression analysis [Tasker and Stedinger 1989; Griffis and Stedinger, 2007b].

Veilleux [2009] provided a review of the hydrologic regional regression literature motivating the use of B-GLS for regional regression analysis. For many years, regional regression analysis used an Ordinary Least Squares (OLS) framework that considers the residual errors to be homoscedastic and independently distributed [Riggs, 1973]. Stedinger and Tasker [1985,1986 ab] developed a GLS framework, which considers both differences in record lengths resulting in variations in precision, as well as cross-correlation among station estimators that result from the cross-correlation among concurrent annual maxima flows at two gauge sites. This spatial correlation arises due to the fact that basins in close proximity to one another can

experience their maximum flows from the same hydrologic event, so that the records upon which flow statistics are computed are correlated, resulting in cross-correlated streamflow statistics. Stedinger and Tasker showed that a GLS analysis provides better estimates of the model parameters and the model error variance in terms of mean squared errors than does an OLS approach. [See also Kroll and Stedinger, 1998] The GLS procedure has been widely used nationally and internationally in many hydrologic studies, including the regionalization of flood quantiles, water quality parameters, low-flow statistics, and extreme rainfall [Tasker *et al*., 1986; Curtis, 1987; Tasker and Driver, 1988; Landers and Wilson, 1991; Moss and Tasker, 1991; Ludwig and Tasker, 1993; Rosbjerg and Madsen, 1995; GREHYS, 1996; Madsen and Rosbjerg, 1997; Robson and Reed, 1999; Kjeldsen and Rosbjerg, 2002; Feaster and Tasker, 2002; Madsen *et al*., 2002; Micevski and Kuczera, 2009; Parrett *et al*., 2011]. GLS has also been used as a regression method in various studies using region-of-influence (ROI) techniques to estimate flood quantiles [Tasker *et al*., 1996; Law and Tasker, 2003, Eng *et al*., 2007a, Eng *et al*. 2007b].

Reis *et al*. [2003, 2005] introduced a Bayesian approach to parameter estimation for the Generalized Least Squares (GLS) regression analysis developed by Stedinger and Tasker [1985, 1986ab] for regional hydrologic analysis. A Bayesian analysis [Zellner, 1971; Gelman *et al*., 2004] provides both an exact measure of precision of the model error variance that method of moment (MM) and maximum likelihood (ML) estimators lack, and a more reasonable description of the possible values of the model error variance in cases where the MM and ML model error variance estimators are zero or nearly zero [Madsen and Rosbjerg, 1997]. The results presented in Reis *et al*. [2005] show that for cases in which the model error variance is small compared to the sampling error of the at-site estimates, which is often the case for regionalization of the skewness coefficient, the Bayesian posterior distribution provides a more

reasonable description of the model error variance than both the MM and ML point estimators. The MM estimator of the model error variance can be zero if the observed variability in the data is explained by the sampling error in the at-site estimates, causing a distortion in the uncertainty of the regional estimate. Similarly, the ML estimator of the model error variance may not be a good representation of the possible values of the model error variance when its value is small or zero because the likelihood function is often highly skewed; this results in the mode being a less appropriate summary statistic than the center-of-mass. Sometimes, the mode is at the origin, which results in a ML estimate of zero even through non-zero values are very likely. The Bayesian-GLS regression framework for hydrologic analysis and corresponding diagnostic statistics have since been expanded upon by Veilleux [2009] and Gruber *et* al. [2007]. Bayesian-GLS regression analysis has been used to develop regional skewness models in several locations in the United States, including the Southeastern U.S. [Veilleux, 2009; Feaster *et al*., 2009; Gotvald *et al*., 2009; Weaver *et al*., 2009; Gruber and Stedinger, 2008] and California [Parrett *et al*., 2011].

1.3 Dissertation Organization

Chapter 2 of this dissertation introduces the GLS regional regression framework and then provides a detailed examination of leverage, influence and the misrepresentation of beta variance statistic for GLS regression. Special attention is devoted to new leverage and influence metrics for use in GLS regression analysis. Examples are supplied which illustrate the different types of leverage and influence metrics, and the insight they contribute to a regional regression analysis. An update of the misrepresentation of beta variance (MBV) diagnostic statistic is also presented along with an example comparing the old formulation to this proposed revision. The MBV

diagnostic statistic is used to determine if a WLS regression is sufficient, or if a GLS regression is needed.

Chapter 3 extends the quasi-analytic Bayesian analysis of the Generalized Least Squares (GLS) regional hydrologic regression framework introduced by Reis *et al*. [2005] to more accurately and precisely estimate regional skewness coefficients. Chapter 3 focuses on using on using an extended B-GLS framework to develop a regional skew model for the State of California. The extended framework described in that chapter was developed due to the extremely large cross-correlations among California annual peaks which caused the B-GLS procedure described in Reis *et al*. (2005) to become statistically unstable. Also, prior to performing the regional skewness analysis in California, a low outlier test (Expected Moments Algorithm) was employed with the California annual peak flow records and subsequently those records were adjusted, resulting in modified at-site skewness estimators.

Chapter 4 builds on the extended B-GLS framework presented in Chapter 3, and develops a more general B-WLS/B-GLS framework for regional hydrologic regression analyses. In that analysis, B-WLS/B-GLS framework first uses OLS analysis to generate stable variances of each at-site skewness estimator, then B-WLS is used to generate an estimator of the regional skewness model parameters, and finally B-GLS is used to estimate the precision of the regression parameters and the model. An example of this B-WLS/B-GLS framework is provided using a dataset from the Southeastern U.S. The B-WLS/B-GLS framework has also been used to generate regional models for flood series of different durations in California [Lamontagne *et al*., 2011].

Finally, Chapter 5 describes the accomplishments of this research focusing on regional regression methods. In particular the extended B-WLS/B-GLS regression model is shown to be

an operational regional hydrologic regression methodology. The research documented in this dissertation provides examples that illustrate the performance of the B-WLS/B-GLS analysis in the estimation of regional skewness coefficients. In addition, the discussion and examples provided of leverage and influence metrics illustrate the information they provide in a GLS analysis and demonstrate their usefulness in identifying rogue observations and effectively addressing lack-of-fit.

REFERENCES

Beard, L. R., (1974), *Flood Flow Frequency Techniques*, Center for Research in Water Resources, The University of Texas at Austin.

Curtis, G.W. (1987), Technique for estimating flood-peak discharges and frequencies on rural streams in Illinois, U.S. Geological Survey Water-Resources Investigations Report 87-4207.

Eng, K., Milly, P.C.D., and Tasker, G.D. (2007a), Flood regionalization: A hybrid geographic and predictor-variable region-of-influence regression method, *Journal of Hydrologic Engineering*, v. 12, p. 585 – 591.

Eng, K., Stedinger, J.R., and Gruber, A.M. (2007b), Regionalization of streamflow characteristics for the Gulf-Atlantic Rolling Plains using leverage guided region-of - influence regression, in Kabbes, K.C., ed., Proceedings of the World Environmental and Water Resources Congress, May 15–19, 2007, Tampa, Florida, USA: American Society of Civil Engineers.

Feaster, T.D. and G.D. Tasker (2002), Techniques for Estimating the Magnitude and Frequency of Floods in Rural Basins of South Carolina, 1999, Water Resources Investigations Report 02-4140, U.S. Geological Survey: Columbia, South Carolina.

Feaster, T.D., Gotvald, A.J., and Weaver, J.C., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 3, South Carolina: U.S. Geological Survey

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., (2004), *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, FL.

Gotvald, A.J., Feaster, T.D., and Weaver, J.C., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 1, Georgia: U.S. Geological Survey Scientific Investigations Report 2009–5043, 120 p.

Griffis, V.W., and J. R. Stedinger, (2007a), The LP3 distribution and its application in flood frequency analysis, 2. Parameter Estimation Methods, *J. of Hydrol. Engineering,* 12(5), 492-500.

Griffis, V. W., and J. R. Stedinger, (2007b), The Use of GLS Regression in Regional Hydrologic Analyses, J. of Hydrology, 344(1-2), 82-95, [doi:10.1016/j.jhydrol.2007.06.023].

Griffis, V.W., and J. R. Stedinger, (2007c), Evolution of Flood Frequency Analysis with Bulletin 17 J. of Hydrol. Engineering, Volume 12(3), 283-97.

Griffis, V.W., and J. R. Stedinger, (2009a), Closure: The LP3 distribution and its application in flood frequency analysis, 2. Parameter Estimation Methods, *J. of Hydrol. Engineering 14(2)*, 209-212.

Griffis, V.W., and J. R. Stedinger, (2009b), Log-Pearson Type 3 Distribution and Its Application in Flood Frequency Analysis. III: Sample Skew and Weighted Skew Estimators, *J. of Hydrol. Engineering 14(2)*, 121-120, [doi:10.1061/(ASCE)1084-0699(2009)14:2(121)]

Groupe de recherche en hydrologie statistique (GREHYS) (1996), Presentation and review of some methods for regional flood frequency analysis, J. Hydrol. 186 1–4, pp. 63–84.

Gruber, Andrea M., Dirceu S. Reis Jr., and Jery R. Stedinger, (2007), Models of Regional Skew Based on Bayesian GLS Regression, Paper 40927-3285, World Environmental & Water Resources Conference - Restoring our Natural Habitat, K.C. Kabbes editor, Tampa, Florida, May 15-18.

Gruber, Andrea M. and Jery R. Stedinger, (2008), Models of LP3 Regional Skew, Data Selection and Bayesian GLS Regression, Paper 596, World Environmental and Water Resources Congress – Ahupua'a, Babcock, R.W. and R. Watson editors, Honolulu, Hawai'I, May 12-16.

Hardison, C.H., (1975), Generalized skew coefficients of annual floods in the United States and their application, *Water Resour Res.* 11(6), 851-854.

Interagency Committee on Water Data (IACWD). (1982). *Guidelines for determining flood flow frequency: Bulletin 17B (revised and corrected)*, Hydrol. Subcomm., Washington, D.C., 28.

Kjeldsen, T. R. and D. A. Jones (2009), An exploratory analysis of error components in hydrological regression modeling, Water Resour. Res., 45, W02407, doi:10.1029/2007WR006283.

Kjeldsen, T.R. and D. Rosbjerg (2002), Comparison of regional index flood estimation procedures based on the extreme value type I distribution, Stoch. Env. Res. Risk A., 16(5), 358-373.

Kroll, C.N., and J.R. Stedinger, (1998), Regional hydrologic analysis: Ordinary and generalized least squares revisited, *Water Resour. Res. 34*(1), 121-128.

Lamontagne, J., J. Stedinger, J. Ferris, D. Knifong, A. Veilleux, and D. Curry, (2011), Regional Skews for 1-Day, 3-Day, 7-Day, 15-Day, and 30-Day Duration Discharge for the Central Valley Region of California, Report Series XXXX-XXXX, U.S. Geological Survey (in press).

Landers, M.N. and K.V. Wilson, Jr. (1991), Flood Characteristics of Mississippi Streams, Water Resources Investigations Report 91-4037, U.S. Geological Survey in cooperation with Mississippi State Highway Department, Jackson, Mississippi.

Law, G.S., and G.D. Tasker (2003), Flood-frequency prediction methods for unregulated streams of Tennessee, 2000, U.S. Geological Survey Water Resources Investigations Report 03-4176.

Ludwing, A.H. and G.D. Tasker (1993), Regionalization of Low-Flow Characteristics of Arkansas Streams, U.S. Geological Survey Water-Resources Investigations Report 93-4013.

Madsen, H., P. S. Mikkelsen, D. Rosbjerg, and P. Harremoes (2002), Regional estimation of rainfall intensity-duration-frequency curves using generalized least squares regression of partial duration series statistics, Water Resour. Res., 38(11), 1239, doi:10.1029/2001WR001125.

Madsen, H., and D. Rosbjerg, (1997), Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling, *Water Resour Res., 33*(4), 771-782.

McCuen, R.H., (1979), Map skew ???, *J. Wate Resour. Plan. And Manage. Div.*, ASCE, 105(WR2), 265-277 [with Closure 107(WR2), 582, 1981].

McCuen, R.H., (2001), Generalized flood skew: map versus watershed skew, *J. Hydrologic Eng.*, ASCE, *Vol*. 6(4), 293-299.

Micevski, T. and G. Kuczera (2009), Combining site and regional flood information using a Bayesian Monte Carlo approach, Water Resour. Res., 45, W04405, doi:10.1029/2008WR007173.

Moss, M.E. and G.D. Tasker (1991), An intercomparison of hydrological network-design technologies, Hydrological Sciences Journal, 36(3), 209.

Parrett, C., Veilleux, A., Stedinger, J.R., Barth, N.A., Knifong, D.L., and Ferris, J.C., 2011, Regional skew for California, and flood frequency for selected sites in the Sacramento–San Joaquin River Basin, based on data through water year 2006: U.S. Geological Survey Scientific Investigations Report 2010–5260, 94 p.

Reis, D. S., Jr., J.R. Stedinger, and E.S. Martins, (2003), Bayesian GLS Regression with application to LP3 Regional Skew Estimation, Proceedings World Water & Environmental Resources Congress 2003, Editors P. Bizier and P. DeBarry, Philadelphia, PA, American Society of Civil Engineers, June 23-26.

Reis, D. S., Jr., J. R. Stedinger, and E. S. Martins, (2005), Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation, *Water Resour. Res.*, *41*, W10419, doi:10.1029/2004WR003445.

Riggs, H. C., (1973), Regional Analyses of Streamflow Characteristics: Techniques of Water-Resources Investigations of the United States Geological Survey, Book 4, Chapter B3.

Robson, Alice and Duncan Reed (1999), *Flood Estimation Handbook Volume 3: Statistical procedures for flood frequency estimation*, Institute of Hydrology, Oxfordshire, UK.

Rosbjerg, D. and H. Madsen (1995), Uncertainty measures of regional flood frequency estimators, Journal of Hydrology, 167, 209-224.

Stedinger, J.R. and V.W. Griffis, (2008), Flood Frequency Analysis in the United States: Time to Update. (editorial) *J. of Hydrol. Engineering*, April, pp. 199-204.

Stedinger, J.R., and G.D. Tasker, (1985), Regional Hydrologic Analysis, 1. Ordinary, Weighted and Generalized Least Squares Compared, *Water Resources Research, 21*(9), 1421-1432.

Stedinger, J.R. and G. Tasker, (1986a), Correction to "Regional hydrologic analysis, 1, Ordinary, weighted and generalized least squares compared", *Water Res. Research*, 22(5), 844.

Stedinger, J.R. and G. Tasker, (1986b), Regional hydrologic analysis, 2: Model-error estimators, estimation of sigma and log-Pearson Type 3 distributions, *Water Res. Research*, 22(10), 1487-1499.

Tasker, G.D., (1978), Flood frequency analysis with a generalized skew coefficient, *Water Resources Research*, 14(2), 373-376.

Tasker, G.D. and N.E. Driver (1988), Nationwide Regression Model for Predicting Urban Runoff Water Quality at Unmonitored Sites, Water Resources Bulletin, 24(5), 1091-1101.

Tasker, G.D., J.H. Eychaner and J.R. Stedinger (1986), Application of Generalized Least Squares in Regional Hydrologic Regression Analysis, in Selected Papers in the Hydrological Science, U.S. Geological Survey, Water Resources Division, Reston, VA, Water Supply Paper 2310, pp. 107-116, December.

Tasker, G.D., S.A. Hodge, and C.S. Barks (1996), Region of Influence Regression for Estimating the 50-Year Flood at Ungauges Sites, Water Resources Bulletin, 32(1), 163-170.

Veilleux, A.G. (2009), Bayesian GLS Regression for Regionalization of Hydrologic Statistics, Floods and Bulletin 17 Skew, M.S. Thesis, Cornell University, August.

Weaver, J.C., Feaster, T.D., and Gotvald, A.J., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 2, North Carolina: U.S. Geological Survey Scientific Investigations Report, 2009-5158.

Zellner, A., (1971), *An Introduction to Bayesian Inference in Econometrics*, John Wiley and Sons, Inc., New York.

CHAPTER 2

LEVERAGE, INFLUENCE AND MBV DIAGNOSTIC METRICS FOR GLS REGIONAL REGRESSION FRAMEWORK

2.1 Introduction

This chapter focuses on leverage, influence, and the misrepresentation of beta variance (MBV) as diagnostic metrics for use with the Generalized Least Squares (GLS) framework for regional regression. First, the GLS methodology is developed focusing on its application to regional hydrologic characteristics, such as the coefficient of skewness. That is followed by a discussion of the assumption that for well formulated models, the models errors are independent across sites. Then, several leverage and influence metrics that appear in the literature are discussed and alternatives are developed. Examples illustrate the performance of these metrics with simple examples and real data sets. Finally, problems with the previous definition of the MBV are discussed, a new MBV definition is proposed and then the two are compared with an example.

2.2 Summary of Generalized Least Squares Methodologies

Stedinger and Tasker [1985,1986] and Tasker and Stedinger [1989] developed a Generalized Least Squares (GLS) regression framework for use in hydrologic regression. In their regression framework it is assumed that the actual value of the quantity of interest y_i for a given site i can be described by a function of physiographic characteristics with an additive error

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \delta_i \qquad i = 1,2,\ldots n \text{ sites} \tag{2.1}$$

where X_{ij} (j=1…k) are the elements of a matrix of k explanatory variables based upon the physical characteristics for each site used in the regression model, $\boldsymbol{\beta}$ is an (k x 1) vector of regression parameters, and δ_i are assumed to be independently distributed model errors with mean zero and variance σ_δ^2. However, in most analyses, only an estimate $\hat{y}_i$ of y_i is available, and thus a time-sampling error η_i should be introduced into the model. As formulated by Stedinger and Tasker [1985, 1986] and further developed in Reis *et al.* [2005, eqn. 6], the GLS model becomes

$$\hat{y} = \mathbf{X}\beta + \delta + \eta = \mathbf{X}\beta + \varepsilon \qquad \text{where } \hat{y}_i = y_i + \eta_i \qquad i = 1,2,\ldots n \text{ sites} \tag{2.2}$$

Thus the observed regression model errors ε_i are the sum of the model errors δ_i and the sampling errors η_i. The total error vector ε has mean zero and a covariance matrix

$$E\left[\varepsilon\varepsilon^T\right] = \Lambda(\sigma_\delta^2) = \sigma_\delta^2\mathbf{I} + \Sigma(\hat{\mathbf{y}}) \tag{2.3}$$

where $\Sigma(\hat{\mathbf{y}})$ is the covariance matrix for the sampling errors in the sample estimators $\hat{y}_i$, $\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)$ is the (n x n) GLS covariance matrix of $\boldsymbol{\varepsilon}$, $\hat{\mathbf{y}}$ is the (n x 1) vector of observed data, and n is the number of sites . The generalized least squared estimator of $\boldsymbol{\beta}$ is [Stedinger and Tasker, 1985; eq 11]

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1}\hat{\mathbf{y}} \tag{2.4}$$

where $\mathbf{X}$ is the (n x k) matrix of basin characteristics with a first column of ones and k is the number of regression parameters. This is the minimum variance linear unbiased estimator and has the following covariance [Greene, 2003]

$$Var\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}^T \boldsymbol{\Lambda}\left(\sigma_\delta^2\right)^{-1} \mathbf{X}\right)^{-1} \tag{2.5}$$

The practical estimation of the covariance matrix for quantile regression involves a number of approximations developed in Tasker and Stedinger [1989], whose impact on the precision of the regression is explored by Kroll and Stedinger [1998]. The GLS framework can be used with streamflow data to derive empirical relationships between physical watershed characteristics and hydrologic characteristics for a site, such as the T-year flood, low flow statistics, or the log-space skewness coefficient used to fit a log-Pearson Type III distribution.

Reis *et al.* [2005] developed a Bayesian-GLS (B-GLS) analysis of the GLS framework, which can provide a better description of the likely value of the model error variance. The B-GLS regression analysis requires specification of prior distributions for both the **β** parameters and the model error variance σ_δ^2. An almost non-informative multivariate normal distribution with a mean of zero and a large variance is used as the prior for β. An exponential distribution with a single parameter λ was used as the prior for the model error variance σ_δ^2. The λ parameter is the reciprocal of the prior mean for σ_δ^2. Reis *et al.* [2005], set λ to 6 with the expectation that as experience accumulates a larger value of λ, corresponding to a smaller prior mean, could be justified. After determining the prior distributions, Reis *et al.* [2005] calculated the posterior moments of the **β** parameters and the full posterior distribution of the model error variance σ_δ^2. In doing so, they showed that B-GLS provides a more realistic description of possible values of the model error variance, especially in cases where the sampling error variances are larger than the model error variance. Gruber *et al.* [2007] and Veilleux [2009] provided a more detailed discussion of the B-GLS methodology including leverage, influence, Pseudo R_δ^2 and pseudo ANOVA.

A critical step, when using the above GLS (or B-GLS) methodology to perform a regional regression on skew, is estimating the cross correlation between skewness coefficient estimators $\hat{y}_i$ and $\hat{y}_j$ for two sites *i* and *j*. To develop the best possible estimate of the regional skew estimator and to understand its precision, it is important to represent the cross correlation between skewness coefficient estimators as accurately as possible.

Martins and Stedinger [2002] used Monte Carlo experiments to determine the needed relationship as a function of the cross correlation of concurrent annual maximum flows ρ_{ij}. Their cross correlation model is

$$\rho\left(\hat{y}_i, \hat{y}_j\right) = Sign\left(\hat{\rho}_{ij}\right) cf_{ij} \left|\hat{\rho}_{ij}\right|^{\kappa} \quad \text{where } cf_{ij} = n_{ij} \Big/ \sqrt{\left(n_{ij} + n_i\right)\left(n_{ij} + n_j\right)} \tag{2.6}$$

wherein n_{ij} is the common record period, n_i and n_j are the extra observation periods and κ is a constant between 2.8 and 3.3. As shown in Equation 2.6, the second factor cf_{ij} included in the model accounts for the sample size difference between the site, as well as the concurrent record lengths. In Tasker and Stedinger [1989], Reis *et al.* [2005], and later Gruber *et al.* [2007], the inter-site correlation coefficient between concurrent flows ρ_{ij} is modeled solely as a function of the distance between two gauge sites.

After developing models using the methodology described in Equations 2.1-2.6, diagnostic statistics are needed to determine which of several candidate models provides the best fit. Such diagnostic statistics are developed in the next section below. Descriptive statistics have been developed to evaluate how well the model describes the data. The goal of model selection is to resolve which set of possible explanatory variables best fit the data affording the most accurate prediction, while also allowing for the simplest model possible. Leverage (Section 2.4) and influence (Section 2.5) are two descriptive statistics used to evaluate the fit of the regression model to the data, model adequacy and data quality.

2.3 Cross-Correlation Among Model Errors

As described in Section 2.2, the GLS regression framework includes two errors, the time-sampling error η that results from the finite length records, and the model error δ intrinsic to the regression model's lack of perfection. If the cross-correlation among the concurrent floods were modeled accurately, the statistical analysis would capture the cross-correlation among the time-sampling errors η. In general, the analysis should capture approximately the average cross-correlations among flood flows at different sites a given distance apart, and thus should on average provide a reasonable estimate of the cross-correlation among the errors in estimated streamflow statistics as a function of the two record lengths, the length of the concurrent record, and the distance between any two pairs of sites.

As discussed in Veilleux [2009], a more troubling concern is possible correlation among the model errors δ. Such correlation would occur if two sites in the model represent nearly the same hydrologic experience, *i.e.* the two sites physically overlap, and thus, are not independent experiences but rather for the most part are the same watershed. For example, this could occur if the ratio of the drainage areas DA_i/DA_j (where $DA_i > DA_j$) is equal to 1.2 and the drainage basins are nested. The basins are one within the other and differ by only 20% in drainage area; thus they are for the most part the same watershed.

In that case, instead of being two independent spatial observations depicting how drainage basin characteristics are related to a dependent hydrologic variable, these two basins are the instead the same spatial experience. In Section 2.2, such site pairs are referred to as redundant. In such cases, the statistical analysis incorrectly represents the information in the data. In the GLS regression model, each individual equation for each individual site

$$\hat{y}_i = x_i\boldsymbol{\beta} + \delta_i + \eta_i \tag{2.7}$$

is intended to represent a different and unique spatial experience. It is this belief that justifies assuming that the individual δ_i are independently distributed so that the covariance matrix of the model errors is

$$\sum(\delta) = \sigma_\delta^2 \mathbf{I} \tag{2.8}$$

For this assumption to be valid, it is critical that an attempt is made to retain only sites that are different spatial hydrologic experiences. If drainage areas overlap to a large extent, then this assumption is violated and the basins are no longer independent spatial experiences.

It would be possible to try to model the cross-correlation among the δ_i if basins had large overlaps, but this is difficult because δ_i is never observed. Rather, only the total errors, ε_i equal to $\delta_i + \eta_i$ are observed. Given the observed variance of the ε_i, and theoretical variance of η_i, the variance of δ_i can be estimated. However, the specific δ_i are not observed. Given the lack of precision with which the cross-correlations of the η_i are known, it would be very difficult to resolve the cross-correlation among the δ_i.

In addition to redundant sites, interdependence among the δ_i could arise in another way. For example, suppose that in the western United States, the mean flood, the coefficient of variation of floods, or the log-space skew were observed to vary with basin elevation. Then if one built a simple GLS regression model that did not include this important explanatory variable, there would be unexplained signal related to location, and thus there would again be cross-correlation among basins that were near each other and thus likely to have similar elevations. In such a situation, it must be decided if the dependence upon elevation will be modeled with explanatory variable in the regression analysis or with an explanatory variable in a model of the cross-correlation among the model errors.

In general modeling such physical dependence of the y-variable on physiographic parameters is much more effectively done by including such variables in the regression. This allows for a direct and immediate understanding of the impact of the variable on the predictive ability of the model, the corresponding precision of the model parameters, and the predicted value of *y*.

Alternatively, the dependence of the y-variable on physiographic parameters can be included in the correlation of spatially dependent model errors. However, this would cause the physiographic parameter (*i.e.* elevation) to be approximated by a spatial relationship. If one has an elevation effect, it seems the best course of action is to incorporate that relationship directly in the regression model so it is employed consistently, rather than by representing that physical elevation-flood relationship as cross-correlation that is explained by distance. Moreover, when there is correlation among the model errors, the prediction at any site should be computed employing the observed residual errors at nearby and related sites so as to incorporate the information provided by that spatial correlation. This could become a very involved task and could prove difficult in locations where there are no neighboring gauge sites to illustrate the effect of the physiographic variable (*i.e.* elevation). In this case the effect of the physiographic variable would be neglected in predictions. Clearly the wise decision is to include any signal provided by physiographic variables in the regression, rather than leaving it to a model of the spatial correlation of the model errors. However, looking for cross-correlation among the model errors, which is explained by distance, is a good way to check if any physiographic or climate variable has been overlooked.

Robson and Reed [1999] and Kjeldsen and Jones [2006, 2009] adapt the GLS regression analysis outlined in Section 2.2 for the estimation of hydrologic variables to include possible

cross-correlation between the regression model errors, δ in Equation 2.7. In particular, Robson and Reed [1999] make the assumption that the cross-correlation among the model errors are the same as the cross-correlation among concurrent floods. This assumption makes the analysis easier because the cross-correlation of annual peaks can be estimated using the observed annual peaks at different sites (this analysis has been part of GLS regional flood-quantile modeling efforts since 1985). However, Kjeldsen and Jones [2006] note that there is little reason to believe that assumption is true; cross-correlation among concurrent annual peaks at different gauge sites reflect the size of storms, whereas possible cross-correlation among the model errors are not likely to arise for this same reason. (See below for possible sources of cross-correlation among model errors.) While Kjeldsen and Jones [2006] identified the weakness in this assumption, it wasn't until their later work [Kjeldsen and Jones, 2007 and 2009] that they strove to estimate the spatial cross-correlation of the model errors from the computed model error estimators.

Kjeldsen and Jones [2009] hypothesize that these model errors result from one of two things: i) the correct set of basin characteristics which explain the dependent hydrologic variable is unknown or ii) from errors present in the measurement of those basin characteristics. As noted above, a third reason is that for some pairs of basins, the watershed for the larger basin included the watershed of the smaller site (redundant sites). This is different from the hypothesis made by Stedinger and Tasker [1985] and applied to Tasker and Stedinger [1989] in which the model errors are assumed to be normal and independently distributed with a mean of zero and a constant variance of σ_{δ}^{2}. It is important to note that Kjeldsen and Jones [2009] also assume a constant model error variance.

Kjeldsen and Jones [2009] propose a recursive-bootstrapping GLS procedure to parameterize these cross-correlations between the model errors with an exponential decay model as a function of distance between basin centroids. The first step develops the sampling error covariance matrix, as well as initializes a guess for the model error variance and forms the model error covariance matrix as an identity matrix. Next, an initial "traditional" GLS analysis is performed, in which the cross-correlation among the sampling errors is ignored. From this analysis the regression parameters and the 'raw' regression residuals are estimated. Following that initial GLS analysis, the raw residuals are reweighted and then used to generate an updated estimate of the model error variance. Then, the updated model error variance estimate is used to generate an updated set of residuals. These updated residuals are then used to estimate the correlation function of the model errors. This process is repeated until a specified tolerance is reach on the estimate of the model error variance.

In the recursive-bootstrapping GLS procedure described above, Kjeldsen and Jones [2009] use a maximum likelihood procedure to derive parameter estimates of their regional model including the parameters of the spatial correlation function of the model errors, which is described by the weighted sum of two exponential functions.

A clear reason for seeing cross-correlation among model errors could be due to redundant sites, defined above as two gauge sites which physically overlap and thus represent the same hydrologic experience [Veilleux, 2009]. This could account for the trend Kjeldsen and Jones [2009] find in which the model errors are related to distance between basin centroids. Instead of trying to introduce a correlation structure to address the cross-correlation among model errors of nested basins it seems this issue could be handled more easily by either deleting one gauge site

from each pair of redundant sites or modeling the network structure of the watersheds as suggested by Kjeldsen and Jones [2009, pg 3], but did not applied.

The second reason for seeing cross-correlation among model errors could be due to not having the correct set of explanatory variables (basin characteristics) with which to model the dependent hydrologic variable [Kjeldsen and Jones, 2009]. However, if there is cross-correlation of model errors with basins in close proximity due to a systematic trend in basin characteristics (*i.e.* elevation, soil type or climate) then it seems it would be better to include a description of that trend in basin characteristics as an explanatory variable. Adding a correlation structure to the model errors to address this lack of adequate explanatory variables has the disadvantages discussed above.

As an alternative to either GLS or the recursive-bootstrapping GLS procedure, Renard [2011] proposes a sophisticated hierarchical Bayesian analysis as a framework for regional hydrologic analysis. The framework assumes that concurrent rainfall or flood series have a multivariate copula distribution that allows for spatial cross-correlation which depends upon the distance between any two sites. Generally, previous hierarchical Bayesian analyses [Kuczera, 1983; Cooley *et al.*, 2007; Coles and Casson, 1998] assume that hydrologic series observed at different sites are independent. However, as this is generally not true, the extension of a hierarchical Bayesian analysis described by Renard [2011] is significant. Renard [2011], following Kjeldsen and Jones [2009] assumes that the model errors have a spatial structure and via a MCMC Baysian analysis Renard [2011] fits the cross-correlation function proposed by Kjeldsen and Jones [2009]. In Renard's [2011] analysis of rainfall series in Southern France, elevation was an important explanatory variable, and when elevation is omitted from an analysis of mean precipitation the resulting residuals exhibit significant spatial correlation.

Renard [2011] suggests modeling simultaneously the location and scale parameters for rainfall distributions at each site, as well as modeling only the location parameter. His analysis found that there is cross-correlation between the scale-model-errors and the location-model-errors of nearly 0.50. Modeling the spatial distribution of scale-model errors, location-model errors, and the cross-correlation between the two was not attempted. Such problems could be avoided by focusing the analysis on quantiles rather than on a location and a scale parameter. Quantile regression was the approach deliberately taken by Tasker and Stedinger [1989] to avoid this problem. Renard's [2011] example illustrates well the problems of prediction when using a model with spatially cross-correlated model errors. In this case, the best prediction is no longer given by use of the derived relationship between the mean of at-site parameters and physiographic parameters. Instead the conditional mean of the at-site parameters should be employed, given the model errors associated with observations that are nearby and assuming that there are enough nearby sites to represent the spatial relationship which has not been captured by the x-variables.

This analysis strongly concludes that it is best to include physiographic and climatological information in a regional analysis through the set of x-variables included in the regression model, rather than to expand the analysis by adding a model of spatial correlation among the model errors. If spatial relationships are included in a spatial correlation function, then it is no longer clear what predictive model has been adopted, nor is its precision explicit; both the model adopted and its precision will vary form site-to-site depending upon what neighboring sites are available. Still, the analyst is warned that when considering incomplete regression models that omit important relationships, the model errors are likely to be spatially correlated resulting in a distortion of the computed precision of estimated parameters; this would

be particularly true for index variables that have values of one for some regions and zero elsewhere because the precision of such variables is more sensitive to cross-correlation than are slope parameters [Stedinger and Tasker, 1985].

2.4 Leverage for OLS and GLS Regression

Belsley *et al*. [1980], Cook and Weisberg [1982, pp. 11] and Hoaglin [1988] explain that leverage maps the observed vector of $\hat{y}$ values into the vector of fitted (or predicted) $\tilde{y}$ values. Thus, leverage can identify those sensitive points in the analysis where $\hat{y}_i$ has a large impact on the fit of $\tilde{y}_i$ [Hoaglin and Welsch, 1978]. Generally, leverage considers whether an observation, or x-value, is unusual, and thus likely to have a large effect on the estimated regression coefficients and predictions. If all the residuals have the same units and precision, then this is a reasonable measure of the effect of a unit error in the different observations. Thus, this leverage measures the marginal impact of the residuals ε_i on the estimated y_i-values. Belsley *et al*. [1980] and Cook and Weisberg [1982] define leverage for ordinary least squares (OLS) analysis as

$$h_{ii} = \frac{\partial \tilde{\mathbf{y}}_{OLS,i}}{\partial \varepsilon_i} \tag{2.9}$$

where $\tilde{\mathbf{y}}_{OLS}$ is an (n x 1) vector of the results predicted by an OLS analysis, and h_{ii} are the diagonal elements of the **H** matrix defined in Equation 2.8.

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \tag{2.10}$$

where **X** is an (n x k) matrix of basin characteristics. This follows from the observations that the sample estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{y}}$ and the estimators of the value of **y** at each site can be computed as $\tilde{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

How to measure leverage in a GLS framework can be problematic. In particular, it is not clear how to describe how large a change in different residuals should be considered when model errors are heteroscedastic. The leverage measure suggested by Tasker and Stedinger [1989, eqn. 23], considers the effect of a unit change in each residual and define leverage for a generalized least squares analysis (GLS) to be

$$h_{ii}^{*} = \frac{\partial \tilde{y}_{GLS,i}}{\partial \varepsilon_i} \tag{2.11}$$

where $\tilde{\mathbf{y}}_{GLS}$ is an (n x 1) vector of the y-values predicted by an GLS analysis; h_{ii}^{*} are the diagonal elements of the **H*** as defined as

$$\mathbf{H}^{*} = \mathbf{X}\left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \tag{2.12}$$

where $\boldsymbol{\Lambda}_{GLS}^{-1}$ is the inverse of the (n x n) GLS covariance matrix. Equation 2.11 is analogous to Equation 2.9 for an OLS analysis, assuming that it is appropriate to consider a unit change equal to 1 in each of the residuals. A new statistic, statistical leverage, is introduced at the end of this section to address possible differences in the variance of each ε_i.

De Gruttola *et al*. [1987] consider a three-step GLS analyses for multivariate linear models with repeated measurements. Repeated measurements allow the use of residuals from a step-1 OLS analysis to be used in a step 2 that estimates the covariance matrix $\boldsymbol{\Lambda}$ for the residual errors ε_i and ε_j associated with each measurement. The three-step GLS analyses as defined by De Gruttola *et al*. [1987] can be described as

$$\hat{\boldsymbol{\beta}}_{r,OLS} = \left(\mathbf{X}_r^T \mathbf{X}_r\right)^{-1} \mathbf{X}_r^T \hat{\mathbf{y}}_r \tag{2.13a}$$

$$\hat{\boldsymbol{\Lambda}} = \sum_{i=1}^{n} \varepsilon_{0,i} \varepsilon_{0,i}^T \Big/ n \tag{2.13b}$$

$$\hat{\boldsymbol{\beta}}_r = \left(\sum_{i=1}^{n} \mathbf{X}_{r,i}^T \boldsymbol{\Lambda}^{-1} \mathbf{X}_{r,i} \right)^{-1} \left(\sum_{i=1}^{n} \mathbf{X}_{r,i}^T \boldsymbol{\Lambda}^{-1} \hat{\mathbf{y}}_{r,i} \right) \tag{2.13c}$$

where $\mathbf{X}_r$ is the (*pn* x *k*) matrix of covariates, $\hat{\mathbf{y}}_r$ is the (*pn* x 1) vector of responses, $\varepsilon_{0,i}$ is the vector of residuals from the OLS regression for the *i*th subject, *p* is the number of covariates, *n* is the number of subjects, and *k* is the number of regression parameters. For the third step of their analysis, the GLS regression, they proposed a measure of leverage for their GLS analysis based on Equation 2.11 which Tasker and Stedinger [1989] observe matches their proposal for leverage in Equation 2.11.

Martin [1992] discusses leverage, influence and residuals for data whose errors are a correlated time series. With the traditional definition of leverage in Equation 2.12, observations on the boundaries, for both temporal data and spatial data, will tend to be the points with higher leverage. Martin [1992] defines a scaled complimentary leverage, **Q**, as a generalization of leverage for dependent data, where he defines **Q** as

$$\mathbf{Q} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1} \tag{2.14}$$

Here the diagonal of **Q** is the complementary leverage, **V** is the (*n* x *n*) correlation matrix, $\mathbf{V}^{-1}$ is the (*n* x *n*) inverse correlation matrix **V**, **X** is (*n* x *k*) matrix of covariates, *n* is the number of observations, and *k* is the number of regression parameters. In the notation presented in this section, $\boldsymbol{\Lambda} = \sigma^2 \mathbf{V}$, where σ^2 is the variance of every observation in a time series. Martin [1992] also proposes a scaled complimentary leverage, $q_i = Q_{ii}/v^{ii}$ where Q_{ii} is the diagonal component of **Q,** the complementary leverage, and v^{ii} is the i^{th} diagonal component of the (*n* x *n*) inverse correlation matrix $\mathbf{V}^{-1}$.

Martin [1992] found that when there is a positive dependence among temporal data, and the regression includes a constant, the smallest values of complimentary leverage tend to occur at

the beginning and the end of the sequence. Martin's complementary leverage is different than the leverage metric proposed by De Gruttola *et al*. [1987] and Tasker and Stedinger [1989]. As Martin [1992] discusses, traditional GLS leverage described in Equation 2.12, assigns its smallest values to those observations which are closest to the average of the dataset. However, the complementary leverage proposed by Martin [1992] is counterintuitive as the points with the largest potential impact on the regression have the smallest leverage value. Moreover, with the leverage defined in Equations 2.10 and 2.12, small leverage values have a lower bound of zero; with Martin's complementary leverage defined in Equation 2.14, the upper bound on complementary leverage is not clear. However, the scaled complementary results in an upper bound of 1. The example in Section 2.8.1 shows that Martin's [1992] leverage does not correctly assess the trend between increased correlation and leverage. Thus the metric defined in Equation 2.11 appears to be the better metric to describe leverage.

Another measure of leverage, introduced by Reis [2005], is Statistical Leverage (S-leverage). Statistical Leverage considers not a unit change in each residual, but a change proportional to the standard deviation of that residual. Thus, this measure considers the likely statistical variation in each ε_i and the effect of such variation on the precision of the estimated model. Thus it addresses the problem of heteroscedasticity, wherein the different equations might even have different units; for example, in a groundwater problem, different equations could correspond to head, flow and water surface elevation. S-leverage for the i^{th} observation is defined to be

$$\text{S-leverage}(i) = f \cdot k \cdot \frac{\partial \hat{y}_i}{\partial \varepsilon_i} \cdot \sigma_{\varepsilon_i} = f \cdot k \cdot h_{ii}^{*} \cdot \lambda_{ii} \quad \text{,where } f = 1 \Big/ \left(\sum_{j=1}^{n} h_{ii}^{*} \cdot \lambda_{jj}^{1/2} \right) \tag{2.15}$$

where h_{ii}^{*} are the diagonal elements of the **H*** matrix defined in Equation 2.12, λ_{ii} are the diagonal elements of the (n x n) $\mathbf{\Lambda}_{GLS}$ covariance matrix, and k are the number of regression parameters. As it is defined in Equation 2.15 with the normalization constant f include in the definition, the average value of S-leverage is also equal to k/n. Twice the average value, $2k/n$, is considered to be a large value. As described by Reis [2005] and Veilleux [2009], S-leverage is an appropriate statistic to consider when the concern is with the likely effect on the regression of probabilistic variation in each residual, or when the observations are measured in different units. The GLS weights depend upon the statistical precision of each ε_i. Thus, the leverage in Equation 2.12 for a point often increases as the at-site record length increases because of the greater weights assigned to the observation; S-leverage in Equation 2.15 is less dependent on record length because the standard deviation of each ε_i decreases with the length of record for each site. If an observation has no leverage, then given its anticipated statistical precision and the leverage associated with the corresponding **x**, the observation is unlikely to have any effect on estimated model parameters. The leverage in Equation 2.12 may be more appropriate when one is concerned with the impact of gross errors in a model's structure, but it does not correct for differences in units among the ε_i.

As discussed by Loader [1999, p. 27 and 36], leverage and influence can also be used in the context of locally weighted least squares regression (often called LOWESS or LOESS, See Cleveland, [1979]; Cleveland and Devlin, [1988]). In this case, because a unique regression defines the estimate of y_i at each point x_i, the leverage assigned to the every point in the dataset can be computed for each i. Loader [1999] defines the leverage for each point i as the coefficient on that y-value when estimating the mean response for the function with an x-value of x_i. (Loader

[1999] actually refers to this measure as influence. However in the order to be consistent with the other studies described above, his influence will be referred to as leverage.) Unfortunately, this neglects what should be the more interesting statistics, which are the coefficients on the y_j values for other points x_j, where $j \neq i$, when predicting the response for x_i. The Loader-leverage on point i should be well behaved: LOESS uses a tri-cubed weight function so that x_i is in the middle to the extent possible of the effective region-of-influence that is created. Clearly one result of the tri-cube weight function used with LOESS regression is to down-weight the more extreme observations; such weighting should decrease the leverage for those extreme points. What would be more informative for each i is the largest positive and the most negative coefficients on any point, or the largest squared coefficient on any point (noting that the variance of a prediction depends on the sum-of-squared values of those coefficients, Loader, 1999, eqn. 2-13).

2.5 Influence for OLS and GLS Regression

Unlike leverage which highlights points which have the ability or potential to affect the fit of the regression, influence attempts to describe those points which do have an unusual impact on the regression analysis [Belsley *et al*., 1980; Cook and Weisberg, 1982; Stedinger and Tasker, 1989]. An influential observation is one with an unusually large residual that has a disproportionate effect on the fitted regression relationships. Influential observations often have high leverage. Cook [1977] discusses that in order to determine the degree of influence that the i^{th} observation has on the estimate of the regression parameters, that point could be deleted from the regression and the regression parameters should be re-estimated. Then, the influence would be a measure of the difference between the estimated regression parameters both with and

without the i^{th} observation. That deletion influence is calculated by the following influence measure, D_i, called Cook's D for OLS regression,

$$\begin{aligned} D_{OLS,i} &= \frac{1}{k}\frac{Var\left(\tilde{y}_{OLS,i}\right)}{Var\left(\hat{\varepsilon}_i\right)}\frac{\hat{\varepsilon}_i^2}{Var\left(\hat{\varepsilon}_i\right)} \\ &= \frac{h_{ii}\hat{\varepsilon}_i^2}{k\left(1-h_{ii}\right)^2\hat{\sigma}_\varepsilon^2} \end{aligned} \tag{2.16}$$

where $Var\left(\hat{\varepsilon}_i\right) = \left(1-h_{ii}\right)\hat{\sigma}_\varepsilon^2$ as derived by Cook [1977]. Cook and Weisberg [1982] observe that the ratio of $Var\left(\tilde{y}_{OLS,i}\right)\Big/Var\left(\hat{\varepsilon}_i\right)$ measures the total change in the variance of prediction at observations 1 through *n* when observation *i* is deleted. Hoaglin [1988] states that the deletion approach to calculating influence is powerful because as shown in Equation 2.16, these values can be calculated without actually removing each observation and redoing the regression *n* times. Hoaglin [1988, eqn. 17] proposes the DFITS metric to measure influence of observation *i* on the fitted $\tilde{y}_i$ which is very similar to D_{OLS} in Equation 2.16. The DFITS metric presented in Hoaglin [1988, eqn. 11], proposed by Belsley *et al*. [1980] and amended by Velleman and Welsch [1981] is calculated as

$$DFITS_i = \frac{h_i^{1/2}\varepsilon_i}{\left(1-h_i\right)s^i} \tag{2.17}$$

where h_i is the i^{th} diagonal element of the leverage matrix **H** defined in Equation 2.10, ε_i is the residual for the i^{th} observation, s^i is the estimated error variance when the i^{th} row of **X** and $\hat{\mathbf{y}}$ have been deleted. There are two main differences. First, DFITS is not scaled by $1/k$ and second the square roots of the remaining terms in Equation 2.16 are used. Thus, while measuring the same quantities, DFITS allows its influence measure to be negative as a result of not squaring the residuals.

Another metric used to measure influence is DFBETAS, which focuses on the influence of the i^{th} observation on the regression parameter [Belsley *et al*. 1980; Hoaglin, 1988]. In order to calculate DFBETAS, the difference between the estimated regression parameters both with and without the i^{th} observation would be divided by the standard error the estimated regression parameters [Hoaglin, 1988].

$$DFBETAS_{ij} = \left(\frac{\left[\left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \right]_{ji}}{\sqrt{\sum_{k=1}^{n} \left[\left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \right]_{jk}^2}} \right) \left(\frac{\varepsilon_i}{(1-h_i)s^i} \right) \tag{2.18}$$

where $DFBETAS_{ij}$ is the measure of change in regression coefficient β_j when the i^{th} observation is deleted, h_i is the i^{th} diagonal element of the leverage matrix **H** defined in Equation 2.10, ε_i is the residual for the i^{th} observation, and s^i is the estimated error variance when the i^{th} row of **X** and $\hat{\mathbf{y}}$ have been deleted.

Tasker and Stedinger [1989] extended the OLS influence metric to their GLS analysis by replacing $Var\left(\hat{\varepsilon}_i\right)$ and $Var\left(\tilde{y}_{OLS,i}\right) / Var\left(\hat{\varepsilon}_i\right)$ in Equation 2.16 by the corresponding values for a GLS analysis. The following equations show the development of the Tasker and Stedinger [1989] GLS influence metric.

The presentation needs to start with the basic GLS regression model

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.19}$$

where **y** is an (n x 1) vector of the predicted hydrologic variable, X is an (n x k) matrix of basin characteristics at gauged sites, $\boldsymbol{\varepsilon}$ is an (n x 1) vector of regression errors, k is the number of basin characteristics including the constant, and n is the number of gauged sites. The GLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \hat{\mathbf{y}} \tag{2.20}$$

where $\boldsymbol{\Lambda}_{GLS}^{-1}$ is the inverse of the (n x n) GLS covariance matrix for gauged sites and $\hat{\mathbf{y}}$ is an (n x 1) vector of the observed dependent variable.

The residuals of the GLS regression are estimated as the difference between the observed values and the predicted values,

$$\hat{\varepsilon} = \hat{\mathbf{y}} - \tilde{\mathbf{y}} \tag{2.21}$$

Substituting Equation 2.19 into Equation 2.21 yields

$$\hat{\boldsymbol{\varepsilon}} = \hat{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{2.22}$$

Thus

$$\begin{aligned} \hat{\boldsymbol{\varepsilon}} &= \hat{\mathbf{y}} - \mathbf{X}\left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \hat{\mathbf{y}} \\ \hat{\boldsymbol{\varepsilon}} &= \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1}\right] \hat{\mathbf{y}} \end{aligned} \tag{2.23}$$

From Equation 2.12, $\mathbf{H}^* = \mathbf{X}\left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1}$ and thus

$$\hat{\boldsymbol{\varepsilon}} = \left[\mathbf{I} - \mathbf{H}^*\right] \hat{\mathbf{y}} \tag{2.24}$$

Substituting Equation 2.19 into Equation 2.24 yields

$$\hat{\boldsymbol{\varepsilon}} = \left[\mathbf{I} - \mathbf{H}^*\right]\left[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right] = \left[\mathbf{I} - \mathbf{H}^*\right] \boldsymbol{\varepsilon} \tag{2.25}$$

In order to determine the variance of the estimated residual vector, given $E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right] = \boldsymbol{\Lambda}_{GLS}$, consider

$$Var\left[\hat{\boldsymbol{\varepsilon}}\right] = E\left[\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}^T\right] = E\left\{\left[\mathbf{I} - \mathbf{H}^*\right] \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \left[\mathbf{I} - \mathbf{H}^{*T}\right]\right\} = \left[\mathbf{I} - \mathbf{H}^*\right] \boldsymbol{\Lambda}_{GLS} \tag{2.26}$$

This can be rearranged to obtain

$$Var\left[\hat{\boldsymbol{\varepsilon}}\right] = \left[\boldsymbol{\Lambda}_{GLS} - \mathbf{H}^* \boldsymbol{\Lambda}_{GLS}\right] \tag{2.27}$$

A new matrix **H'** can be defined as

$$\mathbf{H}' = \mathbf{H}^* \boldsymbol{\Lambda}_{GLS} = \mathbf{X}^T \left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \tag{2.28}$$

By substituting **H'** into Equation 2.27, the $Var\left[\hat{\boldsymbol{\varepsilon}}\right]$ can be expressed as

$$Var\left[\hat{\boldsymbol{\varepsilon}}\right] = \left[\boldsymbol{\Lambda}_{GLS} - \mathbf{H}' \right] \tag{2.29}$$

Thus, the variance of the residual at observation *i* can be written

$$Var\left[\hat{\boldsymbol{\varepsilon}}_i\right] = \left[\lambda_{ii} - h'_{ii} \right] \tag{2.30}$$

where h'_{ii} are the diagonal elements of the matrix **H**' and λ_{ii} are the diagonal elements of $\boldsymbol{\Lambda}_{GLS}$ the (*n* x *n*) GLS covariance matrix. Thus, using the $Var\left[\hat{\boldsymbol{\varepsilon}}\right]$ calculated in Equation 2.29, Tasker and Stedinger [1989] proposed the following influence metric for GLS

$$D_{GLS,i} = \frac{1}{k} \frac{Var\left(\tilde{y}_{GLS,i}\right)}{Var\left(\hat{\varepsilon}_i\right)} \frac{\hat{\varepsilon}_i^2}{Var\left(\hat{\varepsilon}_i\right)} = \frac{h'_{ii} \hat{\varepsilon}_i^2}{k\left(\lambda_{ii} - h'_{ii}\right)^2} \tag{2.31}$$

The influence metric in Equation 2.31, utilizes the same ratios as the OLS influence metric in Equation 2.16. Again it is dimensionless, and primarily depends upon the square of the residuals $\hat{\boldsymbol{\varepsilon}}$ times the leverage for each site as measured by $h'_{ii} / \left(\lambda_{ii} - h'_{ii}\right)$ The factor *k* is only for scaling. In most cases $\left(\lambda_{ii} - h'_{ii}\right)$ is essentially λ_{ii}.

De Gruttola *et al*. [1987] develop and compare two influence measures for a GLS multivariate linear regression with repeated measurements. Critical steps in their analysis were use of an initial OLS regression to estimate the residual errors, and then the use of the residuals for different subjects to estimate the covariance among the set of observations available for each subject. The third set is the GLS regression itself. They note that the influence of an observation

on the estimated regression parameter should be measured including the impact from each of the three steps; the OLS estimate of the regression parameters $\boldsymbol{\beta}$, the estimation of the covariance matrix Λ, and the GLS regression using that estimated covariance matrix, which does not correspond to any of the steps in the GLS analysis in Equations 2.1-2.6 above. Their first influence metric measures the marginal change in the estimated parameter when the weights on the observations are perturbed. This derivate influence, developed by De Gruttola *et al*. [1987] considers the influence of the i^{th} observation of each subject on the entire regression. The influence value developed by Tasker and Stedinger [1989] does not address repeated measurements and only considers the third step, the GLS regression using the estimated covariance matrix. Thus, the influence metric developed by Tasker and Stedinger [1989] focuses on the influence of measurement *i*, on the observation. Their influence metric ignores the impact of each residual on the estimated model error variance, while also considering the entire covariance matrix to be fixed.

The second metric proposed by De Gruttola *et al*. [1987] is deletion influence, which the authors define as the change in the parameter estimate resulting from dropping a set of measurements, which could be all the measurements for one subject from the regression. The authors point out that this method can be computationally intensive and incomplete data sets can be created due to the deletion of a single observation which would require special estimation techniques.

Haslett and Hayes [1998] discuss two complementary types of residuals associated with general linear models (including GLS) with correlated errors. The first type is the marginal residual, which is the classical residual calculated as $\hat{\mathbf{e}} = \hat{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\beta}}$ where $\hat{\mathbf{y}}$ is a vector of the observed values, $\mathbf{X}$ is a matrix of the data and $\hat{\boldsymbol{\beta}}$ is a vector of the estimated regression

coefficients. These marginal residuals are the residuals used by Tasker and Stedinger [1989], Cook [1977] and Cook and Weisburg [1982] to calculate the influence metric. Haslett and Hayes [1998] characterize these marginal residuals as those which measure deviations from the global aspects of the fitted model. The second type of residual presented by Haslett and Hayes [1998] is the conditional residual. This residual is characterized by leaving-out a subset of the data and fitting the model to the remaining data. The conditional residual measures more local aspects by highlighting those marginal residuals which have large conditional correlations. In the case of the regional regressions presented here, the subset could be a single observation. Thus the conditional residual could be considered as the value ε_i of residual i from its conditional mean given the values of other residuals with which it was correlated. While a raw residual may be large, given that other residuals with which it is highly correlated are also large, the value of ε_i can be considered reasonable or expected, and not a value that should cause alarm even though the studentized raw residual is markedly different from zero.

Zewotir and Galpin [2005] focus on creating computationally inexpensive diagnostics to evaluate linear mixed models. A linear mixed model include both fixed effects and random effects, such that

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon} \tag{2.32}$$

where **Y** is a vector of the observed quantity, **X** and **Z** are specified matrices describing the data and the error structure, **β** are the traditional fixed effects to be estimated by the regression, and **u** are the random effects that have a specified distribution [*i.e.* $N\left(0, \sigma_i^2 \mathbf{I}\right)$], and ε are the regression model errors distributed $N\left(0, \sigma_\varepsilon^2 \mathbf{I}\right)$. In particular, they develop four influence metrics all based on Cook's D. These metrics consider the impact of dropping each observation on i) the variance

components ratio $\sigma_i^2 / \sigma_\varepsilon^2$, ii) on the coefficients for the fixed effects, iii) on the value of the random effects, iv) on the likelihood function, and v) on the prediction **Y**. That impact is then expressed as a dimensionless Cook's statistic, just as the influence measure for GLS developed by Tasker and Stedinger [1989] is a dimensionless Cook's statistic. Below a measure of the influence an observation has in the Tasker-Stedinger GLS model on the estimated model error variance is introduced.

Gruber *et al*., [2007] propose another measure of influence, σ-influence which is also described in Veilleux [2009]. It describes, if any, observations have an unusual impact on the estimated model error variance. In using regional skew models, the model error variance is very important because it determines the weight placed on the regional skew relative to the at-site estimator. The σ-influence statistic describes the relative influence of each observation on the estimated model error variance. For example, an observation in the middle of the data set might have a small leverage, and thus even with a large residual, a small value of influence. However, the large residual could still have a major impact on the estimated model error variance.

The influence statistic $D_{GLS,i}$ described in Equation 2.31 identifies those observations with significant influence on the model predictions. $D_{GLS,i}$ does not necessarily describe whether the point has a significant influence on the estimated model error variance. The σ-influence metric is defined as,

$$\sigma - \text{influence}_i = \frac{2\sum_{j=1}^{n} \hat{\varepsilon}_i \left(\mathbf{\Lambda} \left(\sigma_\delta^2 \right)^{-1} \right)_{ij} \hat{\varepsilon}_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} \hat{\varepsilon}_i \left(\mathbf{\Lambda} \left(\sigma_\delta^2 \right)^{-1} \right)_{ij} \hat{\varepsilon}_j} = \frac{2\hat{\varepsilon}_i \left(\mathbf{\Lambda} \left(\sigma_\delta^2 \right)^{-1} \hat{\boldsymbol{\varepsilon}} \right)_i}{\hat{\boldsymbol{\varepsilon}}^T \mathbf{\Lambda} \left(\sigma_\delta^2 \right)^{-1} \hat{\boldsymbol{\varepsilon}}} \tag{2.33}$$

where $\hat{\varepsilon}_i$ is the residual for observation *i*, $\mathbf{\Lambda}\left(\sigma_\delta^2\right)$ is the (*n* x *n*) GLS covariance matrix using the model error variance, σ_δ^2, and *n* is the number of observations. Here the standardized sum-of-squares $\hat{\varepsilon}^T \Lambda\left(\sigma_\delta^2\right)^{-1} \hat{\varepsilon}$ used to compute the likelihood function for the data, and the generalized method of moments model error variance in Stedinger and Tasker [1985, 1986a], is divided among the *n* different sites. By construction, the average value of σ-influence is 2/*n*, where *n* is the number of sites in the regression; thus, σ-influence values greater than 4/*n* are considered to be large, as is the case with $D_{GLS,i}$. The factor of 2 in the numerator of Equation 2.33 allows σ-influence values to be on the same scale as the influences defined in Equation 2.31.

2.6 Leverage and Influence based on Region-of-Influence Regression

Region-of-Influence (ROI) regression uses a unique region, or set of gauged basins, to predict hydrologic quantities such a flood quintiles at ungauged basins [Burns, 1990; Eng *et al*. 2007a; Eng *et al*. 2007b; Tasker *et al.*, 1996]. These regions of influence are usually comprised of sites which are hydrologically similar in some hydrologic sense to the ungauged basins at which predictions are desired [Eng *et al*., 2007a]. However, there are many approaches to defining hydrologically similar regions. Eng *et al*. [2007a] compare three such approaches: predictor-variable, geographic, and a hybrid region of influence which combines the previous two approaches. Locally weighted least squares (LOESS), as developed by Cleveland [1979], is a sophisticated extension of region-of-influence regression. LOESS weights the information available so that data closest to the point of interest have a weight of one and remote points have weights of zero. Regardless of the approach chosen to define the region of influence, diagnostic

metrics can be applied to the regression results to help identify problems. Leverage and influence metrics specifically for ROI regression are developed below.

2.6.1 x_0- Leverage

For a Region-of-Influence (ROI) analysis using OLS or GLS regression, the predicted dependent quantity at $\tilde{y}_0$ is calculated as,

$$\tilde{\mathbf{y}}_0 = \mathbf{x}_0 \hat{\beta} \tag{2.34}$$

where and $\tilde{y}_0$ is an (n x 1) vector of the observed dependent quantity, $\mathbf{x}_0$ is an (1 x k) vector of explanatory variables, and $\hat{\boldsymbol{\beta}}$ are the estimated regression parameters. When using an OLS analysis, $\tilde{y}_0$ can be expressed as

$$\tilde{y}_{0,OLS} = \mathbf{x}_0 \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \hat{\mathbf{y}} \tag{2.35}$$

When using a GLS analysis, $\tilde{y}_0$ can be expressed as

$$\tilde{y}_{0,GLS} = \mathbf{x}_0 \left(\mathbf{X}^T \Lambda_{GLS}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \Lambda_{GLS}^{-1} \hat{\mathbf{y}} \tag{2.36}$$

where $\mathbf{X}$ is an (n x k) matrix of explanatory variables and $\boldsymbol{\Lambda}_{GLS}^{-1}$ is the inverse of the (n x n) GLS covariance matrix.

Thus, following the Cook and Weisberg [1982], approach for OLS leverage, consider a x_0-leverage for an OLS analysis computed as

$$\mathbf{h}_0^{OLS} = \mathbf{x}_0 \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \tag{2.37}$$

where $\mathbf{h}_0^{OLS}$ is an (1 x n) vector of the leverages for each of the n observations on the prediction at $\mathbf{x}_0$.

Similarly, following Tasker and Stedinger [1989] leverage for GLS, Eng *et al*. [2007b] compute x_0-leverage as

$$\mathbf{h}_0^{GLS} = \mathbf{x}_0 \left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \tag{2.38}$$

where $\mathbf{h}_0^{GLS}$ is an (1 x *n*) vector of the leverages for each gauged site on the prediction at the ungauged basin.

Both OLS and GLS x_0-leverage metrics measure the impact on the estimate of y_0 , an observation with characteristics $\mathbf{x}_0$. The average value of this leverage statistics is 1/n (see Appendix A for a proof). Eng *et al*. [2007b] compared ROI results for large leverage values when they were 2, 4, and 8 times the average, and found that 4 times the average resulted in a reasonable threshold 4/n to define unusually large positive leverage points. However, both OLS and GLS x_0-leverage values can be negative, as shown in the examples in Section 2.7 and Section 2.8. Thus, a review of leverage values should be sensitive to both unusually positive and unusually negative leverage values.

2.6.2 x_0-Influence

As discussed in Section 2.5, influence can be used to identify those observations which do have an unusual impact on the regression analysis. In general, the difficulty is in determining what impact to measure; should the focus be on measuring the impact of measurement *i* on the estimated value of y_i associated with x_i, or measuring some average impact across the data set? ROI regression where the interest is in the estimate of y_0 suggests a more straightforward metric for measuring influence when compared to the traditional influence metrics in Equations 2.16 and 2.31.

By substituting Equation 2.37 into Equation 2.35, dependent quantities for $\mathbf{x}_0$ are predicted by an OLS analysis using

$$\tilde{\mathbf{y}}_{0,OLS} = \mathbf{h}_0^{OLS}\hat{\mathbf{y}} \tag{2.39}$$

where $\mathbf{h}_0^{OLS}$ is the x_0-leverage as defined in Equation 2.37. So, the impact on $\tilde{y}_0$ of the estimated residual at each observation $i, \hat{\varepsilon}_i$, is

$$\Delta\tilde{y}_0^{OLS} = \mathbf{h}_0^{OLS}\hat{\varepsilon}_i \tag{2.40}$$

Thus, an influence metric that will correctly order the impact of the estimated errors $\hat{\varepsilon}_i$ on $\tilde{y}_0$ is x_0-influence defined as

$$D_{0,i}^{OLS} = \frac{n\left(h_{0,i}^{OLS}\right)^2 \hat{\varepsilon}_i^2}{\left(1-h_{0,i}^{OLS}\right)\hat{\sigma}_\varepsilon^2} \tag{2.41}$$

The square resolves the problem that some errors increase the value of the predicted value of y_0, while others decrease the predicted value. The denominator in Equation 2.41 corresponds to the variance of $\hat{\varepsilon}_i$ and thus corrects for the fact that $\hat{\varepsilon}_i$ is the estimated residual.

Similarly, by substituting Equation 2.38 into Equation 2.36, dependent quantities for $\mathbf{x}_0$ are predicted by an GLS analysis using

$$\tilde{\mathbf{y}}_{0,GLS} = \mathbf{h}_0^{GLS}\hat{\mathbf{y}} \tag{2.42}$$

So, the impact on $\tilde{y}_0$ of the estimated residual at each observation $i, \hat{\varepsilon}_i$, is

$$\Delta\tilde{y}_0^{GLS} = \mathbf{h}_0^{GLS}\hat{\varepsilon}_i \tag{2.43}$$

Thus, an influence metric that will correctly order the impact of the estimated errors $\hat{\varepsilon}_i$ on $\tilde{y}_0$ is

$$D_{0,i}^{GLS} = \frac{n\left(h_{0,i}^{GLS}\right)^2 \hat{\varepsilon}_i^2}{\left(\lambda_{ii}-h_{ii}^{'}\right)} \tag{2.45}$$

Here $h_{0,i}^{GLS}$ is the leverage of site *i* on the prediction at $\mathbf{x}_0$ as given in Equation 2.38, λ_{ii} is the diagonal element for site *i* of the GLS covariance matrix $\mathbf{\Lambda}_{GLS}$, $h_{ii}^{'}$ is the diagonal element for site *i* of the **H**' matrix in Equation 2.28, and *n* is the number of gauged sites.

This definition of the x_0-influence, $\mathbf{D}_0$, is dimensionless and has an average value of approximately 1/*n*, with a proposed 8/*n* as a critical value. Eng *et al*. [2007b] compared the results for large influence values when they were 4, 8, and 16 times the average, and found that 8 times the average resulted in identification of most large influence points. While the average value of the leverages is 1/n, some values are positive, and some negative, so that the average values of $\mathbf{h}_{0,i}^{GLS}$ can be much larger than $1/n^2$. Eng *et al.* [2007b] proposed an x_0-influence metric based on $h_{ii}^{'}\hat{\varepsilon}_i^2$; however, their influence metric was not dimensionless and also used the absolute value of $h_{0,i}$ to deal with negative leverage values. The x_0-influence proposed in Equation 2.44 resolves both of these issues.

2.7 Comparison of Leverage Metrics for OLS Regression

This section explores the characteristics of different leverage statistics when used with Ordinary Least Squares (OLS) regression. The traditional leverage and the new x_0-leverage are considered. Section 2.7.1 considers the simple case of OLS regression with x-values uniformly distributed along the x-axis, while Section 2.7.2 considers OLS regression with two independent variables.

2.7.1 Univariate OLS Leverage Examples

To compare the tradition leverage to the x_0-leverage, a simple example using OLS regression is considered. A linear OLS regression employs a model of the following form

$$\tilde{\mathbf{y}}_{OLS} = \mathbf{b}_1 + \mathbf{b}_2\mathbf{X} + \boldsymbol{\varepsilon} \tag{2.45}$$

where $\tilde{\mathbf{y}}_{OLS}$ is an (n x 1) vector of the predicted variable, **X** is an (n x 1) matrix of explanatory variables at each observation i = 1, ..., n, $\boldsymbol{\varepsilon}$ is an (n x 1) vector of regression errors where $\varepsilon \sim N\left(0, \sigma^2\right)$ and n is the number of observations.

For an OLS analysis, the traditional leverage given in Equation 2.10 and the x_0-leverage given Equation 2.37 are rewritten below in terms of the univariate regression in Equation 2.45. Thus for a univariate OLS analysis, traditional leverage for x_i is equal to [Hoaglin, 1988]

$$h_i^{OLS} = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\left(n-1\right)S_x^2} \qquad \text{where} \quad S_x^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \Big/ \left(n-1\right) \tag{2.46}$$

where $\bar{x}$ is the mean of the X values, S_x is the standard deviation of the **X** values, and n is the number of observations. Similarly for a univariate OLS analysis, x_0-leverage for at site x_0 from site x_i equals

$$h_{0,i}^{OLS} = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)\left(x_0 - \bar{x}\right)}{\left(n-1\right)S_x^2} \quad \text{where} \; S_x^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \Big/ \left(n-1\right) \tag{2.47}$$

where x_0 is the x value at which the prediction is being made, $\bar{x}$ is the mean of the X values, S_x is the standard deviation of the **X** values, and n is the number of observations. The two expressions in Equations 2.46 and 2.47 are very similar. However the traditional leverage, h_i^{OLS}, is proportional to $\left(x_i - \bar{x}\right)^2$, where as x_0-leverage, $h_{0,i}^{OLS}$, is proportional to $\left(x_i - \bar{x}\right)$.

As shown in both Equation 2.46 and Equation 2.47, characteristics of the y values and the variance of the residuals do not enter into either leverage calculation. However, it is important to observe that the leverage assigned to the most extreme point depends on the distribution of the x values.

Equations 2.46 and 2.47 show explicitly that traditional leverage and x_0-leverage will have the same value when $x_i = x_0$. This follows from the definition of the two leverage statistics being the partial derivative with respect to residual i of the y-prediction made for x equal to either x_i or x_0. When $x_i = x_0$, traditional leverage and x_0-leverage are equal.

Figure 2.1 below compares the traditional leverage (from Equation 2.10 or Equation 2.46) and the x_0-leverage (from Equation 2.37 or Equation 2.47). In Figure 2.1, x values are uniformly distributed and start at a value of -1 and increase by a step size of 0.01 up to a value of +1, thus $n = 21$. Figure 2.1 contains the traditional leverage values for each of the 21 sites in the study. as well the x_0-leverage values for each of the 21 x values for the 21 sites, with $x_o = -1$, $x_o = 0$, and $x_o = +1$, corresponding to predictions being made at those three x_o values. It is important to note that the average value of the traditional leverage is k/n. So in this case with $k = 2$, the average value is $2/n$ or (0.095). The average value of x_0-leverage is $1/n$ or (0.048), which is half as large. However, some x_0-leverage values are negative and some positive.

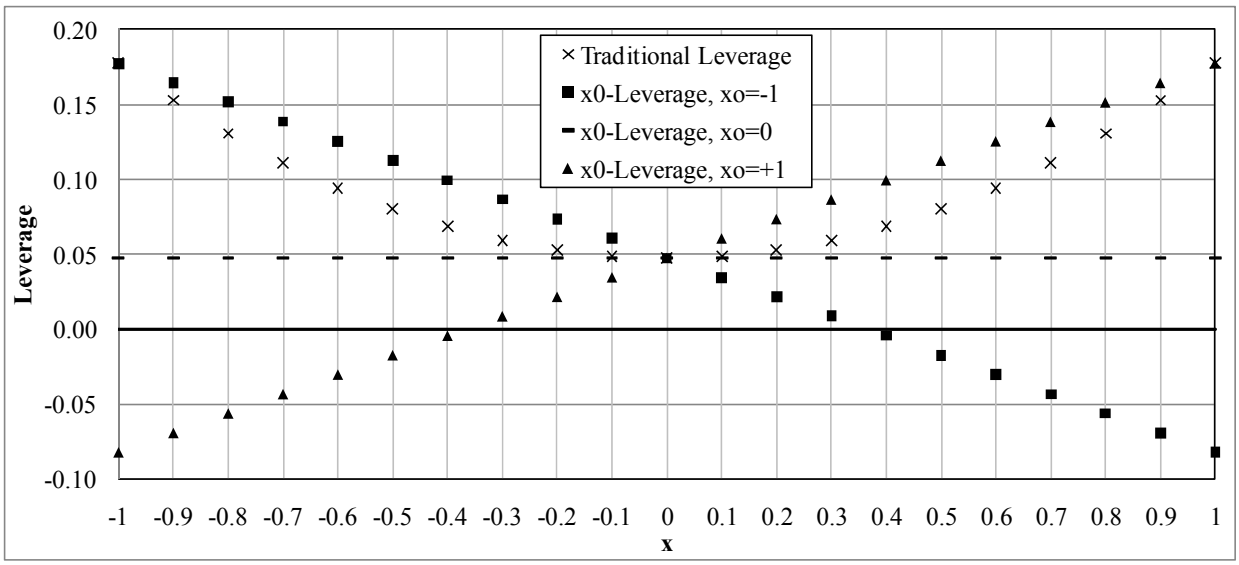


Figure 2.1: Comparison of traditional and x_0-leverage values for an OLS regression. The x's represent traditional leverage h_i^{OLS}. The squares represent x_0-leverage $h_{0,i}^{OLS}$ when predicting at $x_0 = -1$, the dashes represent x_0-leverage when predicting at $x_0 = 0$, and the triangles represent x_0-leverage at $x_0 = +1$.

As shown in Figure 2.1, the traditional leverage and the x_0-leverage can assign very different values to a data point depending upon the value of x_0 selected. The traditional leverage values are independent of the x-values at which future predictions will be made; in fact, the traditional leverage for each i measures the marginal impact of the error in the i^{th} equation on a prediction of y_i obtained with that x_i. Thus with traditional leverage, points with x values at both extremes of the data set (in this example, x values near -1 or +1) have high leverage while the points with average x values (x values near 0) have small leverage. This pattern is very different from the x_0-leverage values.

When predicting a $y(x_0)$-value for a location with a small x_0 value, the leverage values are high for those sites with small x values and decrease as the value of x increases linearly becoming negative at the far extreme. When predicting $y(x_0)$ for a point with a large x_0 value, the leverage values are high for those sites with large x values and decrease linearly as the value

of x decreases, again becoming negative at the extreme. These two lines are the mirror imagines of one another for this symmetric data set. Negative values occur because the predicted y value is a linear combination of the estimated mean value of y plus $\left(x_0 - \bar{x}\right)$ multiplied by a slope term. Estimators of the slope depend on the differences between y values for large x values, and y values for the smaller (negative) x values.

When predicting $y(x_0)$ at a point with a relatively average x_0 value, the x_0-leverage values are relatively constant across the data set; for $x_0 = 0$, the average of the data set, all of the x_0-leverages equal $1/n$. When using local or ROI regression to predict the value of y_0 for a site with the average value of x_0, every site has the same leverage, regardless of the value of x_i. In this case the traditional leverage values are completely misleading.

The traditional leverage metric does not consider the data point where a prediction will be made, or equivalently where the function will be evaluated. The question then is does it make sense to discuss leverage without taking into consideration the characteristics of the location where the function will be evaluated? What does it mean for a site with $x = -0.8$ to have high traditional leverage if the resulting regression model will be used to estimate y at a location with $x_o = +0.8$? It seems that the traditional value of leverage is a measure of the relative density of points in the vicinity of x_i, or equivalently if x_i is in some Euclidean sense different from the other observations. However, whether or not this is good or bad from a statistical viewpoint depends on the value of x_0 at which the function will be evaluated. If the goal is to determine which points in the regression will have the most impact on predicting $y(x_0)$ at a point with x-value x_0, then clearly x_0-leverage should be used because it provides the answer to the question.

A different scenario to consider is one in which the point at which the prediction is to be made has a x_o value outside the limits of data set. For example, if $x_o = +2$. In this case the

leverage for the site with the largest x value would increase drastically. Does this signify that x_o is an unusual site, and thus perhaps a new region of influence needs to be defined that includes other gauged sites more similar to x_o? Or does the high x_0-leverage at the point with $x = 1$, indicate there is something wrong with this point because it has such a large leverage? It seems that the first interpretation is more reasonable: the problem is with x_o and the region, not the observation with the largest x-value. Perhaps limits need to be set on the values of x_o, and if x_o falls outside those limits a new region of influence needs to be defined.

A second OLS regression compares the tradition leverage and the new x_0-leverage when the x values are not symmetric as they were in example above. The x values are generated representing a shifted exponential distribution with a mean of 0. The discrete x_i values are

$$x_i = -\ln\left[\left(n+1-i\right)/n\right] - 0.884 \qquad \text{for } n = 21 \text{ and } i = 1, 2, \ldots, 21 \tag{2.48}$$

The distribution of the x values in Equation 2.48 has a heavy right hand tail and a finite lower bound, as shown in Figure 2.2. Recall that the x values in the first example are uniformly and symmetrically distributed around 0. Figure 2.2 contains the traditional leverage values from Equation 2.10 for each of the 21 sites in the study, as well as the x_0-leverage values from Equation 2.37 when the x_0 values at which prediction are to be made are equal to $x_o = -0.884$, $x_o = 0$, and $x_o = +2.16$ (representing the smallest, the mean, and the largest x-value in the sample). The average value of the traditional leverage is still k/n; in this case $k = 2$, so the average value is $2/n$ or (0.095). The average value of x_0-leverage is again $1/n$ or (0.048).

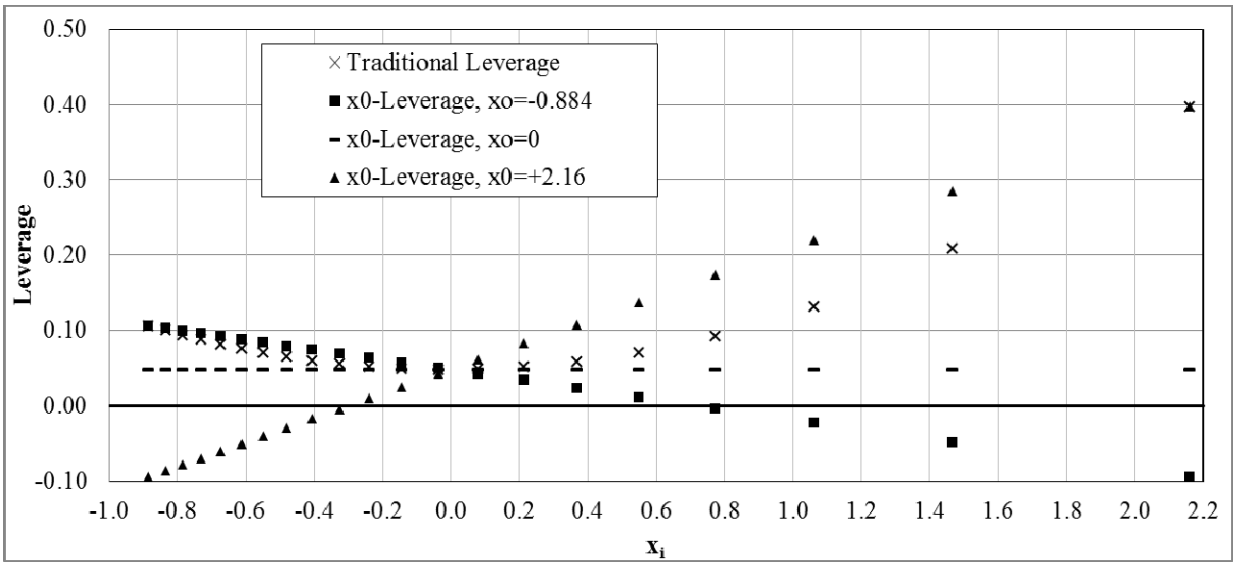


Figure 2.2: Comparison of traditional and x_0-leverage values for an OLS regression with x values drawn from an exponential distribution . The x's represent traditional leverage h_i^{OLS}. The squares represent x_0-leverage $h_{0,i}^{OLS}$ when predicting at x_0 = -0.884, the dashes represent x_0-leverage when predicting at x_0 = 0, and the triangles represent x_0-leverage when predicting at x_0 = +2.16.

Figure 2.2 depicts overall leverage patterns similar to those in Figure 2.1 with several important exceptions. In particular, the leverage functions are no longer symmetric. As shown in Figure 2.2, traditional leverage assigns the highest leverage those with the largest $\left(x-\bar{x}\right)^2$ values, which corresponds to the largest positive x-value in this case has a leverage value eight times the average leverage value; this is due to the asymmetric heavy right-hand tail of the exponential distribution.

Consider the x_0-leverage values for a small x_o value (x_o = -0.884). The site with the smallest x value does not have a large $\left(x_i-\bar{x}\right)$ values and thus even when predicting at x_o = -0.884, the smallest x value in the data set (x = -0.884) does not have a very large leverage value. In this case, the leverages at the smallest x values and the leverages at the largest x values are of opposite sign and almost equal in absolute value. Conversely, when using x_0-leverage to predict

at a site with a large x_o value ($x_o = 2.16$), sites in the region of influence with the most extreme/large x values have leverages which in absolute value are about four times larger than leverage values associated with the small x values in the data set. This is due to the asymmetric heavy right-hand tail of the exponential distribution. When predicting at a site with an average x_o value ($x_o = 0$), the x_0-leverage values all equal $1/n$. A single set of leverage values that is independent of the value of x at which the function will be evaluated fail to provide a telling description.

2.7.2 Bivariate OLS Leverage Examples

The examples in Section 2.7.1 compared traditional leverage and x_0-leverage using a simple univariate dataset. The examples in this section consider leverage metrics for a bivariate regression. Thus, the OLS regression model is assumed to produce a model of the following form

$$\tilde{\mathbf{y}}_{OLS} = \mathbf{b}_0 + \mathbf{b}_1\mathbf{x}_1 + \mathbf{b}_2\mathbf{x}_2 + \boldsymbol{\varepsilon} \tag{2.49}$$

where $\tilde{\mathbf{y}}_{OLS}$ is an (n x 1) vector of the predicted hydrologic variable, $\mathbf{x_1}$ and $\mathbf{x_2}$ are both (n x 1) vectors of basin characteristics at gauged sites, $\boldsymbol{\varepsilon}$ is an (n x 1) vector of regression errors where $\varepsilon \sim N\left(0, \sigma^2\right)$ and n is the number of observations.

For an OLS analysis, the traditional leverage is given in Equation 2.10 and the x_0-leverage is given in Equation 2.37. For these two equations, the $\mathbf{X}$ matrix is an (n x 3) matrix, with a first column of one's for the regression constant, the second column holding the x_1 values from Equation 2.49 and the third column holding the x_2 values from Equation 2.49. Thus, in calculating the x_0-leverage, the x_0 vector is of size (1 x 3) where the first column contains a value

of one and the next two columns hold the values of x_1 and x_2 for the point at which the prediction is being made, respectively.

For the OLS bivariate regression example, consider 21 observations (n =21) with values of $\mathbf{x_1}$ and $\mathbf{x_2}$ symmetrically distributed about (0,0) with $-1.5 \leq x_i \leq +1.5$ for i = 1,2. Figure 2.3 provides a graphic depiction of the ($\mathbf{x_1}$, $\mathbf{x_2}$) pairs.

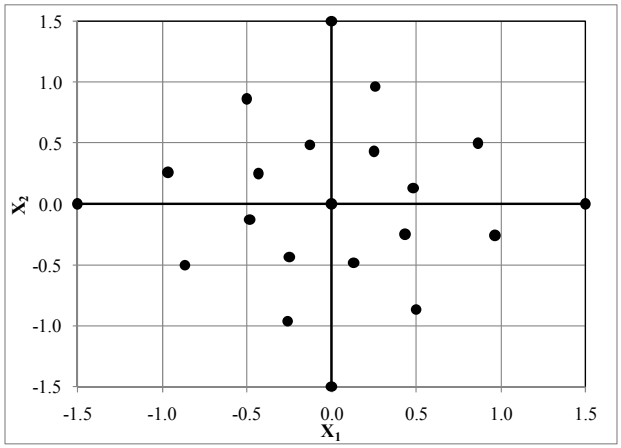
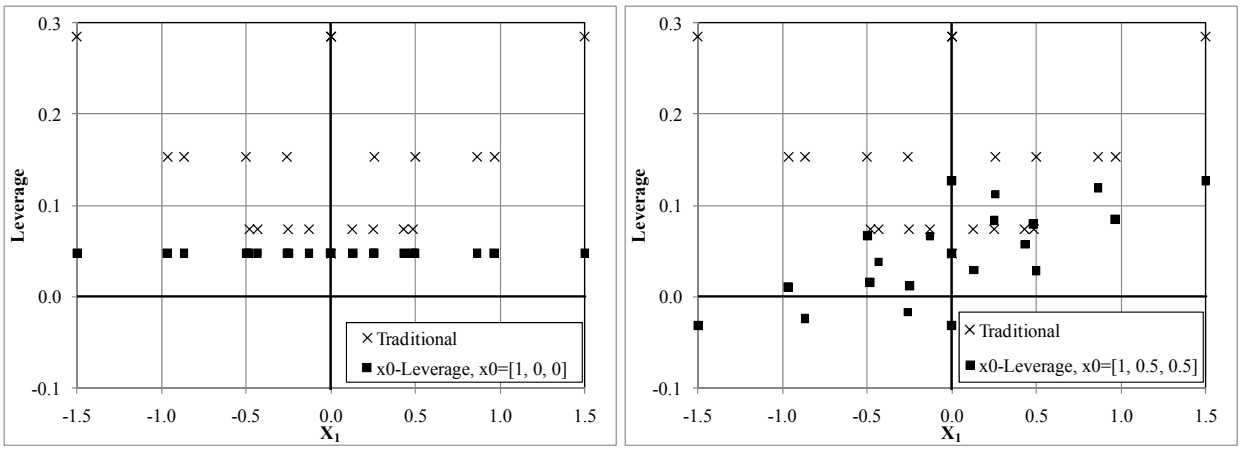


Figure 2.3: Points ($\mathbf{x_1}$,$\mathbf{x_2}$) for symmetric OLS bivariate regression example.

Figure 2.4 contains the traditional leverage values and the x_0-leverage values for each of the 21 observations in the data set. The leverage values are plotted on the y-axis. The x_1 values are plotted on the x-axis in Figures 2.4a and 2.4b. The points were selected so that each unique $x_{1,i}$ value has a unique $x_{2,i}$, except for $x_{1,i}$=0. Figure 2.4a contains the x_0-leverage values when x_o = [1, 0, 0]. Figure 2.4b contains the x_0-leverage values with x_o = [1, 0.5, 0.5].



(a) (b)

Figure 2.4: Traditional and x_0-leverage for an OLS bivariate regression when x_1 and x_2 are symmetrically distributed about the origin. The x's represent traditional leverage and the boxes represent the x_0-leverage. Figure 2.4a contains x_0-leverage values when $\mathbf{x_0}$=[1, 0, 0] and Figure 2.4b contains x_0-leverage values when $\mathbf{x_0}$=[1, 0.5, 0.5].

As Figure 2.4 illustrates, the traditional leverage and the x_0-leverage assign very different values to data points depending upon the value of x_0 selected. The traditional leverage values are independent of the x_0-values at which future predictions will be made. Thus, points with x_1 or x_2 values at the extremes of the data set (x_1 or x_2 values near -1.5 or +1.5) have high leverage while the points with average x_1 or x_2 values (x_1 or x_2 values near 0) have small leverage. The traditional leverage values in Figure 2.4 increase in three steps. These steps correspond to the three concentric data circles in Figure 2.3. The data in the circles closest to [1, 0, 0] have the smallest leverage; the traditional leverage increases as the x values become more extreme. This pattern is very different from the patterns created by the x_0-leverage values.

In Figure 2.4a, the x_0 values at which a future prediction will be made x_0 is a vector with values equal to the center (or average) of the x matrix. Thus, each point is weighted equally in determining the x_0-leverage and each observation has a x_0-leverage value equal to $1/n$; thus no point has unusual leverage because every point gets the same weight. Analogous to the univarite uniform OLS regression, when predicting the value of y_0 for a point with the average

value of x, every point has the same leverage, regardless of the value of x at point *i*. Conversely, the traditional leverage values are not equal; the four most extreme points have leverage values of 0.28, which is twice the average value of 0.14. Thus, in a traditional leverage analysis these four points would be considered high leverage points.

The x_0-leverage results in Figure 2.4b differ drastically from those in Figure 2.4a. In Figure 2.4b, the x_0 values are no longer in the center of the data set. Instead, in Figure 2.3, x_0 lies in the center of the upper right quadrant. Figure 2.4b shows that those observations with positive x_1 values have the largest leverage on the prediction at x_0; observations with negative x_1 values have small and even negative leverage values.

This bivariate OLS regression with symmetric data illustrates the difference between traditional leverage and x_0-leverage. As the traditional leverage metric does not take into account the data point where a prediction will be made, the leverage values at each observation in the study remain the same. This example shows the significant differences in leverage that result from a bivariate analysis. If the goal is to determine which observations in the regression will have the most impact on a prediction y_0 for x-value x_o, x_0-leverage should be used.

A second bivariate OLS regression is considered where the data is not symmetric. Instead, the x_2 values are the same as in the symmetric example above, while the x_1 values from Equation 2.49 are distributed as

$$\mathbf{x}_{1,i} = -\ln\left[\left(n+1-i\right)/21\right] - 0.884 \qquad \text{for } i = 1,2,\ldots 21 \tag{2.50}$$

where *i* is the rank of each $x_{1,i}$ in the symmetric example above. These $x_{1,i}$ values correspond to a shifted exponential distribution with a mean of zero. Figure 2.5 provides a graphic depiction of the 21 (x_1, x_2) pairs.

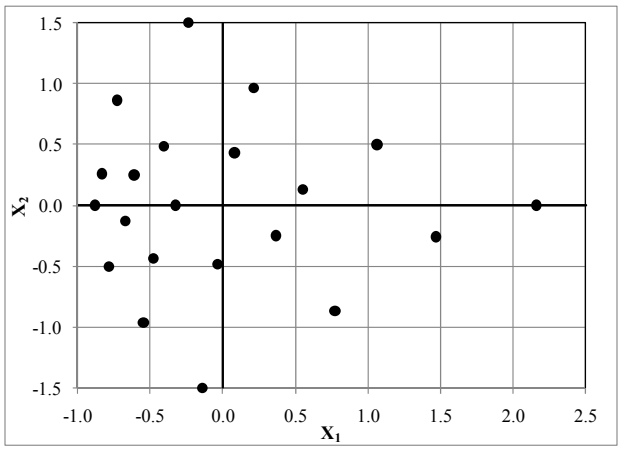
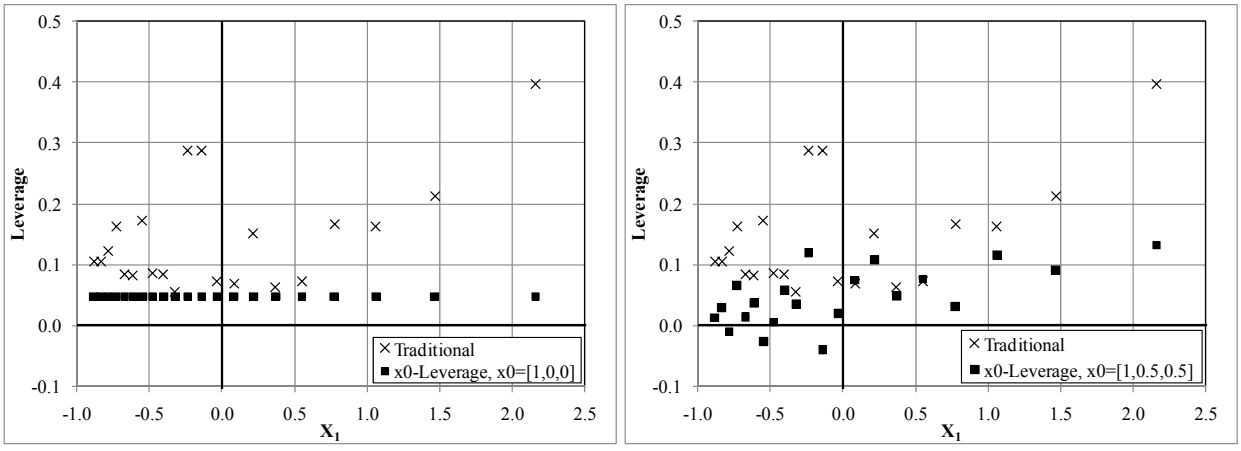


Figure 2.5: Points (x_1,x_2) for shifted exponential OLS bivariate regression example.

Figure 2.6 contains the traditional leverage values and the x_0-leverage values for each of the 21 observations. In Figures 2.6a and 2.6b, the leverage values are plotted on the y-axis. The x_1 values are plotted on the x-axis. The points were selected so that each unique $x_{1,i}$ value has a unique $x_{2,i}$, except for $x_{1,i}$=0. Figure 2.6a contains the x_0-leverage values when x_o = [1, 0, 0]. Figure 2.6b contains the x_0-leverage values x_o = [1, 0.5, 0.5].



(a) (b)

Figure 2.6: Traditional and x_0-leverage for an OLS bivariate regression when x_1 has a shifted exponential distribution with a mean of zero and x_2 is symmetrically distributed about the origin. The x's represent traditional leverage and the boxes represent the x_0-leverage. Figure 2.6a contains x_0-leverage values when x_0=[1, 0, 0] and Figure 2.6b contains x_0-leverage values when x_0=[1, 0.5, 0.5].

Figure 2.6a and Figure 2.6b display the leverage values for this example that employs $x_0 = [1,0,0]$ and $x_0 = [1, 0.5, 0.5]$, respectively. In Figure 2.6a, the x_0-leverage values are all equal to $1/n$; just as in the symmetric OLS bivariate regression example depicted in Figure 2.4a. However, the traditional leverage values show a very large variation, with the point with the most positive and extreme x_1 value having a very large traditional leverage value. The two points with $x_2 = +1.5$ and $x_2 = -1.5$ also have very large leverage values. Again, traditional leverage provides relatively little insight into the actual leverage when predicting at a point. In Figure 2.6b, the traditional leverage values are the same as those in Figure 2.6a. The x_0-leverage values assign the largest leverage values, which are small relative to the large traditional leverage values, to those points with x_1 and x_2 values closest to 0.5; in this case the x_0-leverage value for the smallest x-value is almost zero, rather than having a negative value as it did in Figure 2.4b.

2.8 Comparison of Leverage and Influence for GLS Regression

This section explores the characteristics of different leverage statistics when used with Generalized Least Squares (GLS) regression. Martin's [1992] scaled complementary leverage, traditional leverage and the new x_0-leverage are considered. Section 2.8.1 considers a model with errors that have correlations described by a AR(1) process, while Section 2.8.2 considers leverage and influence metrics resulting from a GLS regression applied to a data set from Illinois River Basin.

2.8.1 Time Series GLS Leverage Example

Martin [1992] considered a model with errors that have correlations described by an AR(1) process. The two examples below consider use of GLS to analyze with temporally

dependent errors. This allows a comparison of traditional GLS leverage in Equation 2.12, x_0-leverage for GLS in Equation 2.38, and the scaled complementary leverage proposed by Martin [1992] reproduced in Equation 2.14.

A linear GLS regression employs a model of the form

$$\tilde{\mathbf{y}}_{GLS} = \mathbf{b}_1 + \mathbf{b}_2\mathbf{x} + \boldsymbol{\varepsilon} \tag{2.51}$$

where $\tilde{\mathbf{y}}_{GLS}$ is an (n x 1) vector of the predicted dependent variable, $\mathbf{X}$ is an (n x 1) matrix of explanatory variables, $\boldsymbol{\varepsilon}$ is an (n x 1) vector of regression errors where $\boldsymbol{\varepsilon} \sim N\left(0, \sigma^2\right)$ and n is the number of observations. The GLS regression parameters, $\mathbf{b}$, are estimated as

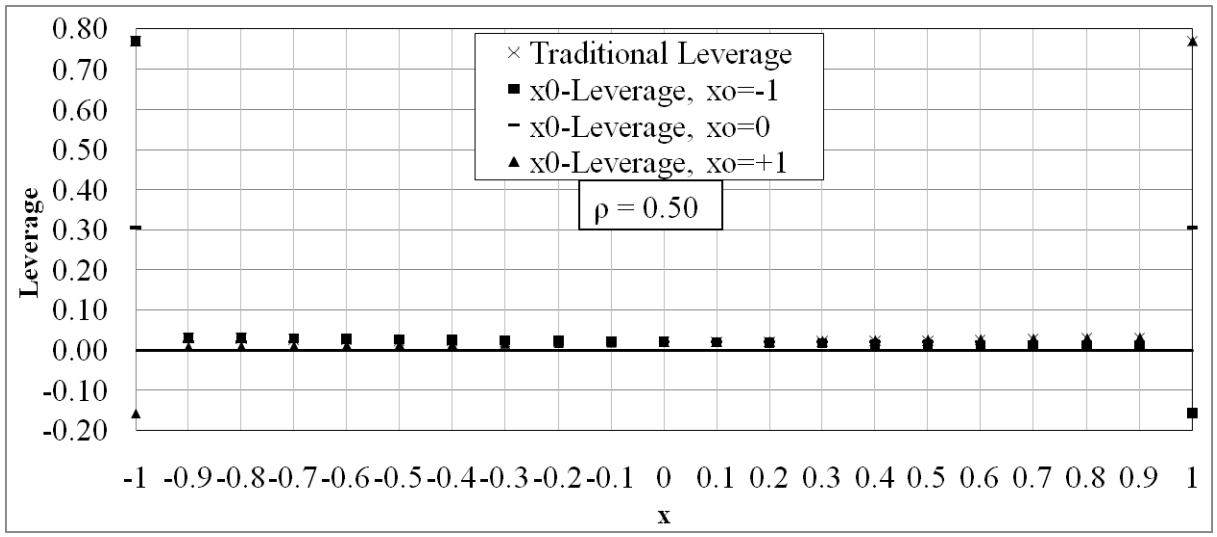
$$\mathbf{b} = \left(\mathbf{X}^T \boldsymbol{\Lambda}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}^{-1} \hat{\mathbf{y}} \tag{2.52}$$

where $\boldsymbol{\Lambda}$ is the (n x n) GLS covariance matrix and $\hat{\mathbf{y}}$ is the observed data. Suppose correlation between two observations depends on their distance apart $|i - j|$, then the covariance matrix $\boldsymbol{\Lambda}$ has the form

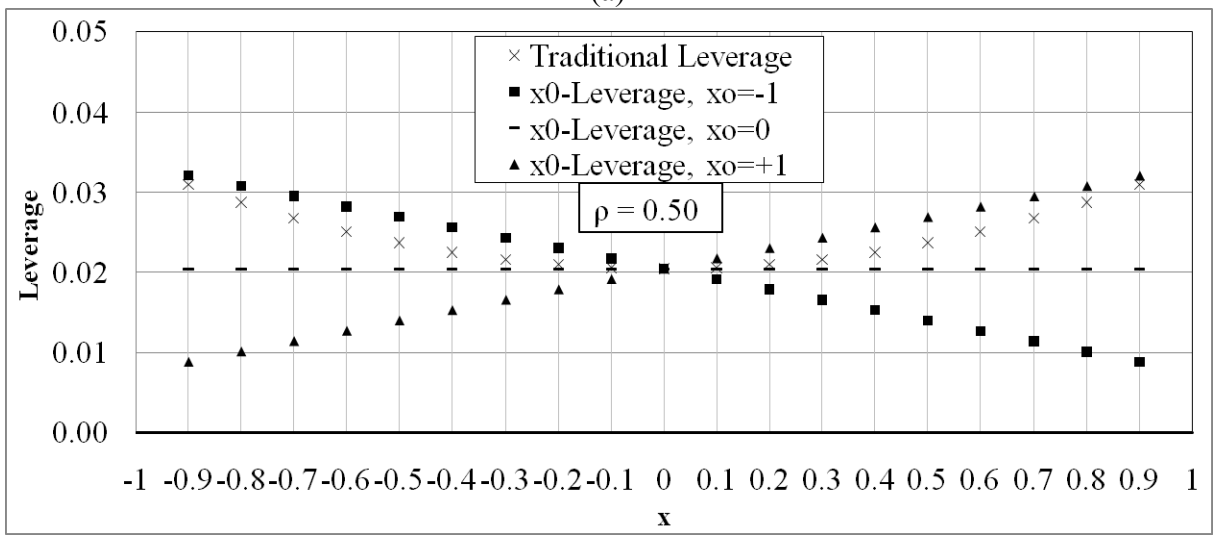
$$\Lambda_{ij} = \rho^{|i-j|}\sigma^2 \quad \text{and} \quad \Lambda_{ii} = 1\left(\sigma^2\right) \qquad \text{where in general } \rho > 0 \tag{2.53}$$

By varying the correlation, ρ, the impact of the correlation on the three different leverage functions can be investigated.

Consider x_i values uniformly distributed between -1 and +1, where again n = 21. Figure 2.7 compares traditional leverage with x_0-leverages (x_o = -1, 0, +1) for a GLS regression with ρ = 0.5 and model error correlation given in Equation 2.53.



(a)



(b)

Figure 2.7: Comparison of traditional and x_0-leverage values for a GLS regression with correlation $\rho = 0.5$. The x-axis of both graphs (a and b) contains the *x* values and the y-axis contains the leverage values. The y-axis in Figure 2.7a ranges from -0.2 up to 0.8. In Figure 2.7b the y-axis is magnified and ranges from 0 to 0.05.

Figures 2.7a and 2.7b plot the leverage versus the x_i values, when $\rho = 0.5$. Figure 2.7a illustrates the leverages at x=-1 and x=+1, however it is difficult to see the pattern in leverage values for the *x* values between those extremes. Thus, Figure 2.7b provides a close up of the leverage values associated with the interior *x* values of the data set by magnifying the y-axis and only showing leverage values between 0 and 0.05.

The traditional leverage values are largest at the extreme x values and are generally the smallest near the average x value. For x_0-leverage when $x_0 = 0$, the leverage follows the same pattern as traditional leverage, however its largest values are less than half as large as the largest traditional leverage values. For x_0-leverage when $x_0 = -1$, the leverage is large and positive when $x = -1$ and small and negative when $x = +1$. The reverse is true for x_0-leverage with $x_o = +1$ leverage, which is large and positive when $x = +1$, and small and negative when $x = -1$.

Figure 2.8 provides five graphs (a-e) which display leverage versus correlation ρ for five different x values, one for each graph; traditional leverage and x_0-leverage with three values of x_0 are included on every graph. The x-axis of all five graphs (a-e) corresponds to the correlation coefficient ρ, and the y-axis corresponds to the leverage values. However, the scale of the y-axis varies from graph to graph; the leverage values for $x = \pm 1$ are much larger than the values obtained with $x = 0$ and $x = \pm 0.5$. When $\rho = 0$, the GLS leverage values are equivalent to OLS leverage values.

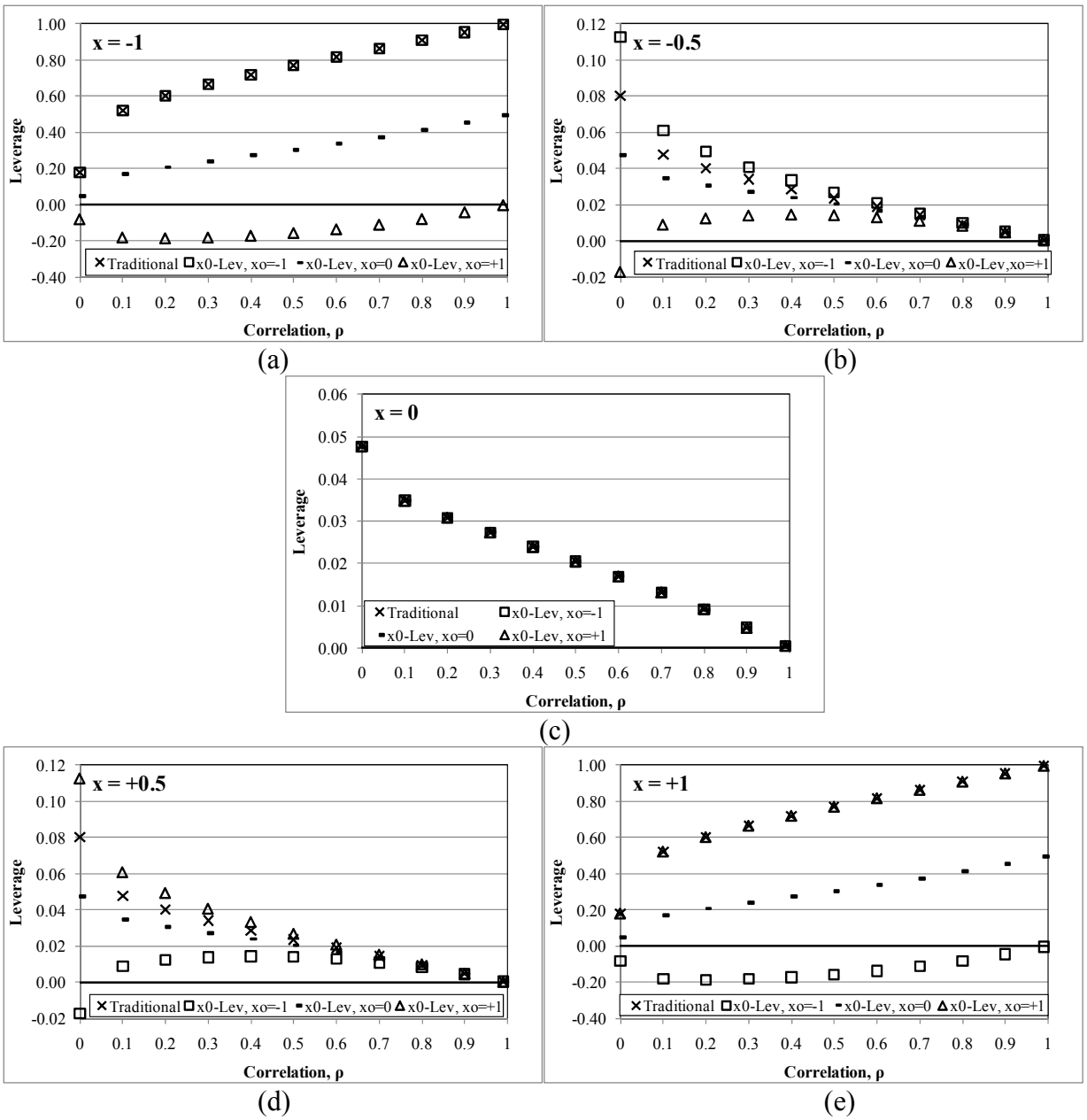


Figure 2.8: Comparison of traditional and x_0-leverage values for a GLS regression with correlation. The five graphs (a-e) plot the leverage versus correlation for a different x values. The x-axis of all five graphs (a-e) contains the correlation coefficient ρ and the y-axis contains the leverage values. The scale of the y-axis varies from graph to graph. In all five figures, the x represent the traditional leverage values, the squares represent the x_0-leverage values when predicting at an ungauged site with $x_0 = -1$, the dashes represent the x_0-leverage values when predicting at an ungauged site with $x_0 = 0$, and the triangles represent the x_0-leverage values when predicting at an ungauged site with $x_0 = +1$.

Figure 2.8a plots the leverage versus correlation when x = -1. The traditional leverage values and the same as the x_0-leverage values when predicting at x_0 = -1, and they reach a maximum leverage of 1 when ρ =1. For all four types of leverage, when the correlation increases the leverage values also increase. Figure 2.8e which plots the leverage versus correlation values when **x** = +1 is exactly the same as Figure 2.8a except that the boxes representing the x_0-levarage: x_0 = -1 leverage and the triangles representing the x_0-leverage: x_0 = +1 leverage are switched. x_0-leverage: x_0= -1 leverage is large and positive when **x** = -1 and small and negative when **x** = +1. The reverse is true for x_0-leverage: x_o = +1 leverage, which is large as positive when **x** = +1 and small and negative when **x** = -1.

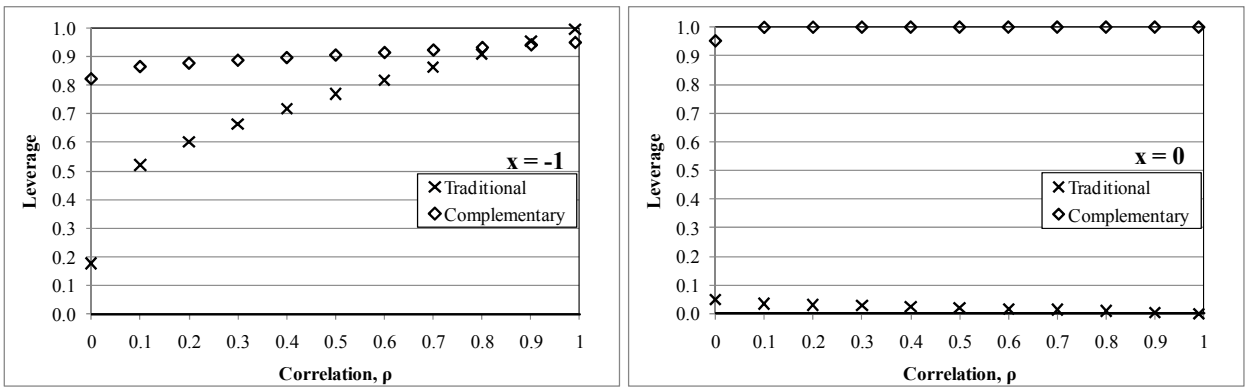
Figure 2.8b displays the leverage versus correlation values when x = -0.5. All four leverage series have zero leverage when ρ =1. The leverage values for traditional leverage, x_0-leverage: x_0=-1, and x_0-leverage: x_0=0, all decrease with increasing correlation for this x-values. x_0-leverage: x_0=+1 has negative leverage when ρ =0, increasing positive correlation up to ρ =0.4, and then decreasing positive correlation until it reaches 0 at ρ =1. Figure 2.8d displays the leverage versus correlation values when x = +0.5 and is exactly the same as Figure 2.3b, except that the boxes representing the x_0-leverage: x_0 = -1 leverage and the triangles representing the x_0-leverage: x_0 = +1 leverage are switched.

Figure 2.8c displays the leverage versus correlation when x =0. For this x-value, four types of leverage decrease with increasing correlation. It is also important to note that all four types of leverage are equal at all values of ρ. At ρ = 0, the leverage is equal to 1/*n* and at ρ =1, the leverage is equal to 0.

Thus the overall trend across all correlation values for traditional leverage versus x, is that leverage is the largest at the two ends where x = -1 and x = +1 and smallest in the middle

where x = 0. The overall trend for x_0-leverage x_o = -1, is that leverage is largest at x = -1 and decreases to x = +1. For x_0-leverage with x_o = +1, is that leverage is the smallest at x =-1 and increases to x = +1. For x_o = 0, when ρ = 0, x_0-leverage is equal to 1/*n* and is constant at all x's. As correlation increases, the x_0-leverage: x_o = 0 at the largest and smallest x's also increases. However, these increases are modest compared to the other leverage metrics.

Figure 2.9 compares traditional leverage with scaled complementary leverage for a GLS regression with correlation as defined in Equation 2.44 representing an AR(1) process. The values of x0-leverage are not included to simplify the graphs. Figure 2.8a displays the leverage for the point x = -1 versus correlation. Figure 2.9b displays the leverage for the point x = 0 versus correlation. The *x*-axis of both graphs represents the correlation coefficient ρ and the *y*-axis represents the leverage values.



(a) (b)

Figure 2.9: Comparison of traditional and complementary leverage values for a GLS regression with correlation. Figure 2.9a plots the leverage for x = -1 versus correlation, while Figure 2.9b plots the leverage for x = 0 versus correlation. The x-axis contains the correlation coefficient ρ and the y-axis contains the leverage values. In both graphs, the x's represent the traditional leverage values and the diamonds represent the complementary leverage values.

As shown in Figure 2.9, traditional leverage and scaled complementary leverage have very different values. It is interesting to note that when ρ=0, the GLS analysis reduces to an OLS analysis and $q_i = 1 - h_{ii}^*$. As Martin [1992] stated, the complementary leverage will be the least at

the beginning and end of the sequence with positive correlation. This is shown to be case in Figures 2.9a and 2.9b for all values of ρ. When comparing the values of complementary leverage in Figure 2.9a (x = -1) with those in Figure 2.9b (x = 0), leverage values in Figure 2.9a are smaller than the corresponding complementary leverage values with the same correlation in Figure 2.9b. Following Martin's [1992] definition of complementary leverage, the x values with the smallest and largest values have the smallest leverage values and thus the smallest impact on the analysis. This is counterintuitive to traditional leverage where the x's with the smallest and largest values have the largest leverage and are considered to have the largest impact on the analysis. As shown in Figure 2.9 as correlation increases, complementary leverage increases and approaches one for both x = -1 and x = 0. Thus, Martin's [1992] complementary leverage indicates that as the observations become more correlated the leverage of all data points approaches one and thus their impact on the analysis becomes more equal. However, traditional leverage correctly observes that as correlation approaches one, all of the leverage moves to the most extreme points. Thus, complementary leverage incorrectly evaluates the relative importance of the points in the data set.

2.8.2 GLS Leverage and Influence Example Using Data from Illinois River Basin

Sixty-two sites from the Illinois River basin were selected to illustrate the use of x_0-leverage and x_0-influence metrics for regional skew regression with a real data set. Record lengths range from 14 to 90 years. This is the same data set that was used in Veilleux [2009] to demonstrate the use of Bayesian Generalized Least Squares (GLS) regression for estimating regional skew. For more details regarding this data set see Veilleux [2009]. The Illinois River basin was divided into three regions, as described in Tasker and Stedinger [1986]. The regions

and values of the binary variables (Z_1, Z_2) for each were: Little Wabash (1, 0), Rock (0, 1), and Sangamon (0, 0). Except for the binary variable, all explanatory variables were centered by subtracting their means so that the constant and the binary variables could be used to compute the regional mean of each hydrologic region.

Table 2.1 contains the results of three Method-of-Moments GLS (MM-GLS) analyses of regional skew based on the Illinois River basin data set. The third model in the table, which includes a constant, a binary region parameter, and the ln(Slope) was the best model as identified in Veilleux [2009]. While that bivariate model provided the best fit, the univariate model, the second model from Table 2.1 will provide a simpler and easier to understand example of the use of x_0-leverage and x_0-influence. Morever, the leverages associated with an index variable such as Z_1 are not particularly interesting because Z_1 takes on only two values (either 0 or 1). This second model contains a constant and ln(Slope). Here ln(Slope) is a statistically significant regression parameter with a p-value of 4% that explains 12% of the variation in the at-site skews.

Table 2.1:MM-GLS regional skew regression results for the Illinois River basin data set with 62 sites. The table reports the AVP_{new}, the average variance of prediction at a site not in the regression, the AVSE, the average sampling variance, σ_δ^2, the model error variance, and R_δ^2 the pseudo R^2 statistic. Below each coefficient is its standard error in parentheses with the p-value reported as a percentage below that.

Model	β_1	β_2	β_3	σ_δ^2	**ASVE**	**AVP_{new}**	R_δ^2
$y = \beta_1$	-0.42 (0.12)	-	-	0.15	0.02	0.17	0%
$y = \beta_1 + \beta_2\left[\ln\left(slope\right)\right]$	-0.31 (0.13)	0.13 (0.06) 4%	-	0.14	0.03	0.16	12%
$y = \beta_1 + \beta_2\left[\ln\left(slope\right)\right] + \beta_3\left[Z_2\right]$	-0.09 (0.17)	0.13 (0.06) 3%	-0.51 (0.24) 4%	0.12	0.04	0.16	24%

Figure 2.10 contains x_0-leverage values and Figure 2.11 contains x_0-influence values for 28 of the 62 sites from the Illinois River basin data set for three possible new sites or equivalently ungauged basins at which regional skewness might be estimated. Those three sites represent small, average, and large slope sites. The 28 sites were chosen because they have either large x_0-leverage or large x_0-influence for one of the three ungauged basins. Thus, to standardize the figures the leverage and influence from same 28 sites are plotted for each of the three ungauged basins. The sites are also sorted so that the sites increase in slope from left to right, (*ie* the site with the smallest slope is the left most, while the site with the largest slope is right most). This allows a visual demonstration of how the x_0-leverage and x_0-influence values at these 28 sites change for each ungauged basin. The residuals for each of the 62 gauged sites are identical in all three cases.

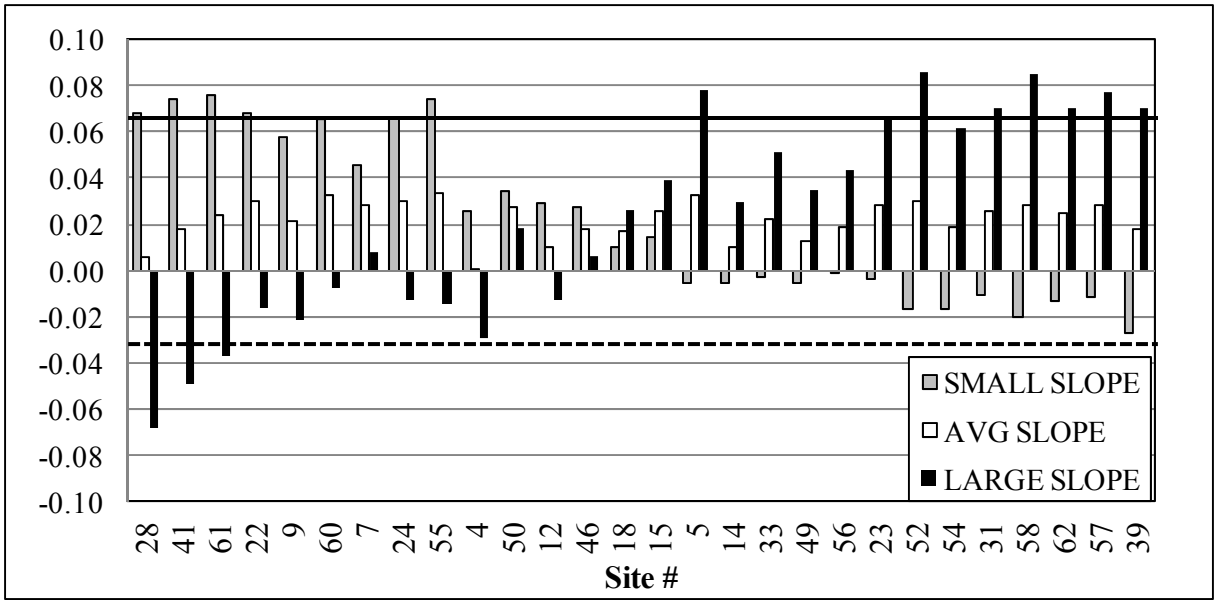


Figure 2.10: x_0-Leverage values for the three ungauged basins. The solid black line represents the threshold for high x_0-leverage while the dashed black line represents the threshold for high negative x_0-leverage. The gray bars represent x_0-leverage for an ungauged site with a small slope, the white bar represent x_0-leverage for an ungauged site with an average slope, and the black bars represent x_0-leverage for an ungauged site with a large slope.

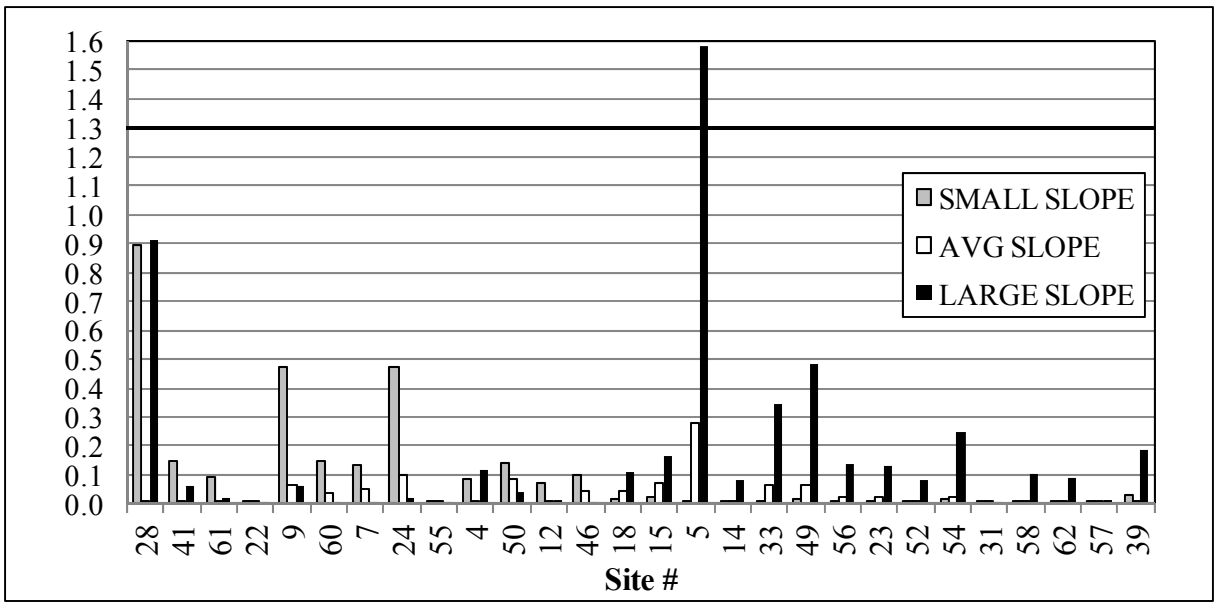


Figure 2.11: x_0-Influence values for the three ungauged basins. The solid black line represents the threshold for high x_0-influence. The gray bars represent x_0-influence for an ungauged site with a small slope, the white bar represent x_0-influence for an ungauged site with an average slope, and the black bars represent x_0-influence for an ungauged site with a large slope.

The three x_0-leverage and x_0-influence metrics differ based on the characteristics of the ungauged site for which the regression is used to predict regional skew. The three "ungauged" sites are chosen to explore the range of slopes present in the data set. Thus, a SMALL SLOPE site was used which has a slope equivalent to Site 28, the site with the smallest slope in the data set. An AVG SLOPE site was used which has a slope equivalent to the average slope in the data set. Finally, a LARGE SLOPE site was used which has a slope equivalent to Site 39, the site with the largest slope in the data set. There are large differences in the x_0-leverage and x_0-influence values for each of the three ungauged basins due to the differences in ln(slope) at each of the three ungauged basins.

A summary of the x_0-leverage and x_0-influence metrics for each of the three ungauged sites in is provided in Table 2.2. The number of sites with high leverage and high influence are tallied. When a site with AVG SLOPE is considered as the site at which the prediction will be

made, there are no sites with either high x_0-influence or high x_0-leverage. However, when either a site with SMALL SLOPE or a site with a LARGE SLOPE are considered as the site at which the prediction will be made, there are gauge sites with high x_0-leverage. Only the LARGE SLOPE site produces sites with high x_0-influence in this example. These results are discussed in greater detail below.

Table 2.2: x_0-Leverage and x_0-Influence results from regional skew regression

Ungauged Site Model	ln(slope)	Centered ln(slope)	# Of Sites With High: Leverage	Influence
SMALL SLOPE	-0.18	-2.35	7	0
AVG SLOPE	2.17	0	0	0
LARGE SLOPE	4.96	2.79	9	1

Figure 2.10 displays the x_0-leverage values for ungauged site with SMALL SLOPE, AVG SLOPE, and LARGE SLOPE. For the AVERAGE SLOPE case there is some modest variation in the leverages due to differences in record length among the sites and cross-correlations. Focusing on the leverages that result when predicting at an ungauged site with SMALL SLOPE (the gray bars), it is evident that those gauged sites with the smallest slope have the largest x_0-leverages. In fact the x_0-leverages for all the gauged sites with large slope have negative x_0-leverages when predicting at a site with SMALL SLOPE. Site 28 has by far the largest x_0-influence on the SMALL SLOPE prediction, which makes sense as it has the smallest slope in the study, centered ln(slope) = -0.17 and a large residual = -0.89.

Comparing the x_0-leverage values for prediction at an ungauged site with SMALL SLOPE to predictions at an ungauged site with a LARGE SLOPE, it is evident that for the most part very different sites impact each prediction. For prediction at an ungauged site with a

LARGE SLOPE, most of the sites with large x_0-leverages are on the right side of the graph, which are the sites with the largest slope. Also, for the most part the sites on the left side of the graph, which are the sites with the smallest slopes, the x_0-leverages are negative when predicting at an ungauged site a LARGE SLOPE. Thus, the x_0-leverage values for SMALL SLOPE and LARGE SLOPE resemble reverse images of one another.

Figure 2.11 displays the x_0-influence values for prediction at an ungauged basin with SMALL SLOPE, AVG SLOPE, and LARGE SLOPE. Unlike x_0-leverage, the x_0-influence values cannot be negative. Thus sites with large x_0-leverages, both positive and negative, may have large x_0-influences depending on the value their residual. While none of the gauged sites have x_0-influence values above the high x_0-influence threshold for prediction at an ungauged site with SMALL SLOPE, it is important to note that those sites with the smallest slopes have the largest x_0-influence values.

Likewise, those gauged sites with large slopes, have on average the largest x_0-influences when predicting at an ungauged site with LARGE SLOPE. However, there is one major exception, Site 28. Because Site 28 has a large negative x_0-leverage (= -0.068) and a large residual (= 0.89), which are both squared in calculating the x_0-influence, it has a large x_0-influence value. Also, when predicting at an ungauged site with LARGE SLOPE, Site 5 is the only gauged site with an x_0-influence value above the threshold. Site 5 has the third largest x_0-leverage (= 0.078) and fourth largest residual of the 62 gauged sites in the study (= 1.3).

Figure 2.10 and Figure 2.11 also display the x_0-leverage and x_0- influence values for prediction at an ungauged basin with AVG SLOPE. None of the gauged sites have large x_0-leverage or large x_0-influence, which is to be expected. Because the ungauged site at which the prediction is occurring is in the middle of the data set, and thus does not represent an unusal case,

none of the leverages should be very unusual. In turn, x_0-leverage will mostly be determined by the size of the residual.

2.9 Misrepresentation of the Beta Variance

When performing a regional statistical analysis, it is important to identify the level of sophistication appropriate for that analysis. At one point in time, ordinary least squares (OLS) was the available tool, and was used for most applications [Riggs, 1973; Stedinger and Tasker, 1985, Tasker and Stedinger, 1986]. Subsequently, weighted least squares (WLS) was adopted for some studies [Tasker, 1980; Stedinger and Tasker, 1985]. Finally, Tasker and Stedinger [1989] introduced an operational generalized least squares (GLS) methodology for regional regression. However at the time the GLS method was developed, simple metrics to guide the selection of the appropriate method were not available.

The Misrepresentation of the Beta Variance (MBV) statistic was developed by Griffis and Stedinger (2007) to evaluate whether a WLS regression is sufficient, or if a GLS regression is needed. The MBV describes the error made by a WLS regression analysis by ignoring cross-correlation among the residuals in its evaluations of the precision of b_0^{WLS}, which is the estimator of the constant β_0. Covariance among the estimated y_i's generally has its greatest impact on the precision of the constant term [Stedinger and Tasker, 1985] and zero-one regional indicator variables, which are constant means for different regions or categories.

Griffis and Stedinger proposed that MBV be measured by the ratio

$$\text{MBV} = \frac{Var\left[b_0^{WLS} \mid \text{GLS analysis} \right]}{Var\left[b_0^{WLS} \mid \text{WLS analysis} \right]} \tag{2.54}$$

The MBV derived by Griffis [2006] and used by Reis *et al*. [2005] and Veilleux [2009] assumes that this WLS problem can be simplified into an OLS problem by scaling each equation with the correct set of weights. Griffis [2006] use weights *m*,

$$m_i = \frac{1}{\sqrt{\lambda_{ii}}} \tag{2.55}$$

where *m* is an (*n* x 1) vector of the inverse of the square root of the variance of the residuals, λ_{ii} are the diagonal elements of **Λ** the (*n* x *n*) GLS covariance matrix, *n* is the number of sites in the regression. Thus, Griffis [2006] calculate the WLS estimator of the regression constant b_0^{WLS}, assuming other regression variables are centered as

$$b_0^{WLS} = \frac{\mathbf{m}^T \hat{\mathbf{y}}}{\mathbf{m}^T \mathbf{v}} \tag{2.56}$$

where **m** is an (*n* x 1) vector of weights as calculated in Equation 2.55, $\hat{\mathbf{y}}$ is an (*n* x 1) vector of the observed data, and **v** is an (*n* x 1) vector of ones. Correspondingly, the variance of the WLS estimator b_0^{WLS} that would be generated by a GLS analysis [Griffis, 2006] is

$$Var\left[b_0^{WLS} \mid \text{GLS analysis} \right] = \frac{\mathbf{m}^T \mathbf{\Lambda m}}{\left(\mathbf{m}^T \mathbf{v} \right)^2} \tag{2.57}$$

Similarly, Griffis [2006] calculates the variance of b_0^{WLS} that would be generated by a WLS analysis as

$$Var\left[b_0^{WLS} \mid \text{WLS analysis} \right] = \frac{\mathbf{m}^T \mathbf{D m}}{\left(\mathbf{m}^T \mathbf{v} \right)^2} = \frac{n}{\left(\mathbf{m}^T \mathbf{v} \right)^2} \tag{2.58}$$

where **D** is an (*n* x *n*) matrix which contains the diagonal components of the (*n* x *n*) **Λ** matrix and contains zeros on the off diagonal. The matrix **D** represents the WLS covariance matrix. Thus, by substituting Equations 2.53 and 2.58 into Equation 2.52, MBV is

$$\text{MBV} = \frac{\mathbf{m}^T \mathbf{\Lambda} \mathbf{m}}{n} \tag{2.59}$$

However, the weights used in the above calculation are not the correct weights for determining the error made by a WLS regression error analysis in its evaluation of the precision of b_0^{WLS}. MBV was computed using weights that are the inverse of the standard deviation. Instead a correct WLS analysis would weight each observation by the inverse of the variance. Thus, a corrected MBV is defined below as MBV* using the correct weights.

The corrected weights *w* are calculated as

$$w_i = \frac{1}{\lambda_{ii}} \tag{2.60}$$

where *w* is an (*n* x 1) vector of the inverse variance of the residuals, λ_{ii} are the diagonal elements of **Λ** the (*n* x *n*) GLS covariance matrix, *n* is the number of sites in the regression. Thus, the WLS estimator of the regression constant b_0^{WLS}, assuming other regression variables are centered is

$$b_0^{WLS} = \frac{\mathbf{w}^T \hat{\mathbf{y}}}{\mathbf{w}^T \mathbf{v}} \tag{2.61}$$

where **w** is an (*n* x 1) vector of weights as calculated in Equation 2.60, $\hat{\mathbf{y}}$ is an (*n* x 1) vector of the observed data, and **v** is an (*n* x 1) vector of ones. Correspondingly, the variance of b_0^{WLS} given a GLS analysis can be calculated as

$$Var\left[b_0^{WLS} \mid \text{GLS analysis} \right] = \frac{\mathbf{w}^T \mathbf{\Lambda} \mathbf{w}}{\left(\mathbf{w}^T \mathbf{v} \right)^2} \tag{2.62}$$

Similarly, the variance of b_0^{WLS} given a WLS analysis can be calculated as

$$Var\left[b_0^{WLS} \mid \text{WLS analysis} \right] = \frac{\mathbf{w}^T \mathbf{D} \mathbf{w}}{\left(\mathbf{w}^T \mathbf{v} \right)^2} = \frac{\mathbf{w}^T \mathbf{v}}{\left(\mathbf{w}^T \mathbf{v} \right)^2} = \frac{1}{\mathbf{w}^T \mathbf{v}} \tag{2.63}$$

where **D** is an (n x n) matrix which contains the diagonal components of the (n x n) **Λ** matrix and contains zeros on the off diagonal. The matrix **D** represents the WLS covariance matrix. Thus, by substituting Equations 2.62 and 2.62 into Equation 2.54, the corrected MBV, MBV*, has a value of

$$\text{MBV}^* = \frac{\mathbf{w}^T \mathbf{\Lambda} \mathbf{w}}{\mathbf{w}^T \mathbf{v}} \tag{2.64}$$

The formula for MBV and MBV* are very similar. Both correctly measure the difference in the variances of an estimator of the model constant that would be provided by WLS and GLS analyses. The difference is that they consider slightly different estimators: MBV considers weights equal to the inverse of the standard deviations, whereas MBV* consider the optimal weights equal to the inverse of the variances. If the variances of the different residuals were all equal, there would be no difference between MBV and MBV*.

2.9.1 Comparison of MBV Values Using Examples from South Carolina and Illinois

The difference between MBV and MBV* from Section 2.9, will be illustrated using the Illinois River basin data set, as well as the South Carolina data set used in Veilleux [2009]. Veilleux [2009] developed regional skew models for these two data sets using Bayesian Generalized Least Squares (B-GLS) regression.

The Illinois River basin data set is comprised of 62 sites with record lengths ranging from 14 to 90 years. The South Carolina data set is comprised of 89 sites with record lengths ranging from 25 to 104 years. All explanatory variables, except binary variables, were centered by

subtracting their means so that the constant and the binary variables could be used to compute the regional mean of each hydrologic region. For more details regarding these data sets see Veilleux [2009].

Table 2.3 shows the MBV and MBV* values for B-GLS regional skew models for both the Illinois River basin data set as well as the South Carolina data set. For each data set, results are provided for both the constant model and the best fit model. (See Veilleux [2009] for detailed regional skew results for these two data sets.) As shown in Table 2.3, the values of MBV and MBV* are almost equal. This indicates that while MBV* is the intended calculation, past results calculated using the old MBV told the intended story: in these cases a GLS analysis is needed to correctly compute the variance of the estimator of the constant in the model. WLS misrepresents the variance of the constant by a factor of about 3 for the Illinois River Basin data set, and by 4.5-5.0 for the South Carolina data set.

Table 2.3: Comparison of MBV and MBV* for Illinois River basin data set (62 sites) and South Carolina data set (89 site) based on B-GLS regional skew models. MBV is calculated according to Equation 2.59 and MBV* is calculated according to Equation 2.64.

Data Set	B-GLS Regional Skew Model	MBV	MBV*
Illinois River Basin	$\tilde{y} = b_0$	2.9	2.8
	$\tilde{y} = b_0 + b_1\left[\ln(\text{Slope})\right]$	3.0	2.9
South Carolina	$\tilde{y} = b_0$	4.5	4.6
	$\tilde{y} = b_0 + b_1\left[\ln(\text{Slope})\right] + b_2\left[\ln(\text{Length})\right]$	4.8	4.9

2.10 Conclusions

A comparison of leverage and influence metrics for use with GLS regression is presented in this chapter. Through derivations and examples, a more clear understanding of how each

leverage and influence metric evaluates the data is provided. Questions are raised regarding the information provided by traditional leverage metrics which do not take into account the *x* value at which the GLS regional regression model will be used to make a prediction. A number of examples demonstrate that x_0-leverage which accounts for the characteristics (or *x*-values) at which a prediction will be made is considered to be a more informative metric. As influence is basically leverage multiplied by the studentized residuals squared, it is important to determine how best to measure leverage in order to develop a useful influence metric. The misrepresentation of beta variance (MBV) statistic is used to determine if a GLS analysis is needed, of if a WLS analysis is sufficient. The work in this chapter proposes that statistic, proposed by Griffis and Stedinger [2007], be revised so that it uses the correct weights. The difference is shown to have relatively little impact on the actual numerical value of the statistic.

APPENDIX A

PROOF FOR AVERAGE VALUE OF x_0-LEVERAGE

This appendix provides a proof detailing that the average value of x_0-leverage is equal to the inverse of the number of observations in the data set.

The GLS regression model is represented as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{A.1}$$

where **y** is an (n x 1) vector of the predicted variable, X is an (*n* x *k*) matrix of explanatory variables for each of the *n* observations, $\boldsymbol{\varepsilon}$ is an (*n* x 1) vector of regression errors, *k* is the number of explanatory variables, and *n* is the number of observations. Thus, the GLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \hat{\mathbf{y}} \tag{A.2}$$

where $\boldsymbol{\Lambda}_{GLS}^{-1}$ is the inverse of the (*n* x *n*) GLS covariance matrix and $\hat{\mathbf{y}}$ is an (*n* x1) vector of the observed dependent variable.

For a regional regression analysis using GLS, the predicted dependent quantity $\tilde{y}_0$ at the point at which the prediction is made is calculated as,

$$\tilde{y}_0 = \mathbf{x}_0 \left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \hat{\mathbf{y}} \tag{A.3}$$

where $\tilde{y}_0$ is the predicted dependent quantity and $\mathbf{x}_0$ is an (1 x *k*) vector of explanatory variables at the point at which the prediction is made. Thus, $\mathbf{x}_0$ is formatted identically to a row in the **X** matrix. It is assumed that $n > k$. The assumption is also made that the values in the first column of the **X** matrix equal 1. The first column of the X matrix will be represented as

$$\mathbf{c}_1 = (1,1,...,1)^T \tag{A.4}$$

where c_1 is a vector with dimensions (n x 1).

x_0-leverages are the partial derivatives of $\tilde{y}_0$ with respect to each component **y**, and thus are the elements of the (1 x n) leverage vector, $\mathbf{h}_0$,

$$\mathbf{h}_0 = \mathbf{x}_0 \left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \tag{A.5}$$

Thus to determine the sum of the leverages

$$\sum_{i=1}^{n} \mathbf{h}_{0,i} = \mathbf{x}_0 \left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{c}_1 = 1 \tag{A.6}$$

Note that the matrix **X** can be written as the combination of k column vectors each of size (n x 1)

$$\mathbf{X} = \left[\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_k \right] \tag{A.7}$$

Thus

$$\begin{aligned} \mathbf{I}(k) &= \left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X} \right)^{-1} \left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X} \right) \\ &= \left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \left[\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_k \right] \end{aligned} \tag{A.8}$$

Where $\mathbf{I}(k)$ is the (k x k) identity matrix. Thus,

$$\left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{c}_1 = \mathbf{e}_1 \tag{A.9}$$

where $\mathbf{e}_1$ is the first unit vector of size (k x 1) and $\mathbf{x}_0 \mathbf{e}_1 = 1$ as required.

REFERENCES

Burns, D.H. [1990], Evaluation of regional flood frequency analysis with a region of influence approach. Water Resour. Res., 26(10), 2257-2265.

Cleveland, W.S. [1979]. "Robust Locally Weighted Regression and Smoothing Scatterplots". Journal of the American Statistical Association 74 (368): 829– 836.doi:10.2307/2286407

Cleveland, W.S. and Devlin, S.J. [1988]. "Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting". Journal of the American Statistical Association 83 (403): 596–610. doi:10.2307/2289282.

Coles, S., and E. Casson [1998], Extreme value modeling of hurricane wind speeds, *Structural Safety*, 20, 283-296.

Cook, R.D. [1977], Detection of Influential Observation in Linear Regression, Technometrics, 19(1), pp. 15-18.

Cook, R.D. and Weisberg, S., [1982], Residuals and Influence in Regression, Chapman and Hall, New York, NY, 230 pp.

Cooley, D., D. Nychka, and P. Naveau [2007], Bayesian spatial modeling of extreme precipitation return levels, *Journal of the American Statistical Association*, 102(479), 824-840.

De Gruttola, Bivrot, James H. Ware, Thomas A. Louis [1987], Influence analysis of generalized least squares estiamtors. Journal of the American Statistical Association, 82(399), pp 911-917.

Eng, K., Milly, P.C.D., and Tasker, G.D., [2007a], Flood regionalization: A hybrid geographic and predictor-variable region-of-influence regression method: Journal of Hydrologic Engineering, v. 12, p. 585 - 591.

Eng, K., Stedinger, J.R., and Gruber, A.M., [2007b], Regionalization of streamflow characteristics for the Gulf-Atlantic Rolling Plains using leverage guided region-of-influence regression, in Kabbes, K.C., ed., Proceedings of the World Environmental and Water Resources Congress, May 15–19, 2007, Tampa, Florida, USA: American Society of Civil Engineers.

Greene, W.H., [2003]. Econometric Analysis: Fifth Edition, Prentice Hall, Upper Saddle River, New Jersey, 1026 pp.

Griffis, V.W., [2006]. Flood Frequency Analysis: Bulletin 17, Regional Information, and Climate Change. Ph.D. Dissertation, Cornell University.

Griffis, V. W., and J. R. Stedinger, [2007b], The Use of GLS Regression in Regional Hydrologic Analyses, J. of Hydrology, 344(1-2), 82-95, [doi:10.1016/j.jhydrol.2007.06.023].

Gruber, Andrea M., Dirceu S. Reis Jr., and Jery R. Stedinger, [2007], Models of Regional Skew Based on Bayesian GLS Regression, Paper 40927-3285, World Environmental & Water Resources Conference - Restoring our Natural Habitat, K.C. Kabbes editor, Tampa, Florida, May 15-18.

Haslett, John and Kevin Hayes [1998], Residuals for the linear model with general covariance structure, Journal of the Royal Statistical Society. Series B (Statistical Methdolody), 60(1), pp. 201-215.

Hoaglin, D.C. [1988], Using Leverage and Influence to Introduce Regression Diagnostics, The College Mathematics Journal, 19(5), pp. 387-401.

Hoaglin, D.C. and Welsch, R.E. [1978], The Hat Matrix in Regression and ANOVA, The American Statistician, 32(1), pp. 17-22.

Kjeldsen, T.R., and D.A. Jones [2006], Prediciton uncertainty in a median-based index flood method using L moments, *Water Resour. Res.* 42, W07414, doi10.1029/2005WR004069.

Kjeldsen, T.R., and D.A. Jones [2007], Estimation of an index flood in the UK, *Hydrol. Sci. J.*, 52, 86-98, doi:10.1623/hysj.52.1.86.

Kjeldsen, T.R., and D.A. Jones [2009], An exploratory analysis of error components in hydrological regression modeling, *Water Resour. Res.* 45, W020407, doi10.1029/2007WR006283.

Kroll, C.N., and J.R. Stedinger, [1998], Regional hydrologic analysis: Ordinary and generalized least squares revisited, *Water Resour. Res. 34*(1), 121-128.

Kuczera, G. [1983], Effect of sampling uncertainty and spatial correlation on an empirical bayes procedure for combining site and regional information, *Journal of Hydrology*, 65, 373-398.

Loader, C. [1999], Local Regression and Likelihood, Springer, New York, NY. 305 pp.

Martin, R. J. [1992], 'Leverage, influence and residuals in regression models when observations are correlated, Communications in Statistics - Theory and Methods, 21(5), pp. 1183-1212.

Martins, E. S., and J.R. Stedinger [2002], Cross correlations among estimators of shape, Water Resour. Res., 38(11), 1252, doi:10.1029/2002WR001589.

Reis Jr., D.S., [2005]. Flood Frequency Analysis Employing Bayesian Regional Regression and Imperfect Historical Information. Ph.D. Dissertation, Cornell University.

Reis, D. S., Jr., J. R. Stedinger, and E. S. Martins, [2005], Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation, *Water Resour.*

Res., *41*, W10419, doi:10.1029/2004WR003445.

Renard, B., [2011]. A Bayesian Hierarchical Approach to Regional Frequency Analysis, *Water Resources Research*, (submitted).

Riggs, H. C., [1973], Regional Analyses of Streamflow Characteristics: Techniques of Water-Resources Investigations of the United States Geological Survey, Book 4, Chapter B3.

Robson, A. and D. Reed, [1999], *Flood Estimation Handbook*, Institute of Hydrology, Wallingford, United Kingdom, Volume 3 Statistical Procedures for Flood Frequency Estimation.

Stedinger, J.R., and G.D. Tasker, [1985], Regional Hydrologic Analysis, 1. Ordinary, Weighted and Generalized Least Squares Compared, *Water Resources Research, 21*(9), 1421-1432.

Stedinger, J.R. and G. Tasker, [1986], Correction to "Regional hydrologic analysis, 1, Ordinary, weighted and generalized least squares compared", Water Res. Research, 22(5), 844.

Tasker, G.D., [1980], "Hydrologic Regression with Weighted Least Squares," Water Resources Research, 16(6), 11107-1113.

Tasker, G.D., and J.R. Stedinger, [1986], Estimating Generalized Skew With Weighted Least Squares Regression, Journal of Water Resources Planning and Management, 112(2), 225-237.

Tasker, G.D., and J.R. Stedinger, [1989], An Operational GLS Model for Hydrologic Regression, Journal of Hydrology, 111(1-4), 361–375.

Tasker, G.D., Hodge, S.A., and Barks, C.S. [1996], Region of influence regression for estimating the 50-year flood at ungauged sites, Water Resour. Bull., 32(1), 163-170.

Veilleux, A. G. [2009], Bayesian GLS Regression for Regionalization of Hydrologic Statistics, Floods and Bulletin 17 Skew, M.S. Thesis, Cornell University, August.

Velleman, P.F. and R.E. Welsch, [1981], Efficient Computing of Regression Diagnostics, The American Statistician, 35, pp. 234-242.

Zewotir, Temesgen and Jacky S. Galpin [2005], Influence diagnostics for linear mixed models. Journal of Data Science, Vol. 3, pp. 153-177

CHAPTER 3

EXTENDED BAYESIAN GLS REGIONAL SKEW ANALYSIS FOR CALIFORNIA ANNUAL MAXIMUM FLOOD FLOWS

3.1 Introduction

An area of active research addresses the development of better statistical procedures to estimate regional skew and its precision. Such advancements in regional skew estimators will improve flood frequency analysis in the United States within the Bulletin 17B framework that employs the log-Pearson Type 3 distribution [Stedinger and Griffis, 2008]. This chapter describes the use of efficient statistical methods to develop a model of regional skew for the State of California, and to estimate its accuracy given the high cross-correlation among concurrent flood peaks at many California sites.

Tasker and Stedinger [1986] developed a Weighted Least Squares (WLS) procedure for estimating regional skew coefficients based on sample skew coefficients corresponding to the logarithms of peak stream flow data. Their regional analysis of skewness estimators accounts for the precision of the skewness estimator for each station, which depends on the length of record for each station, as well as the accuracy of the regional skew model. More recently, Reis *et al*. [2005], Gruber and others [2007], and Gruber and Stedinger [2008] developed a Bayesian GLS regression model. While WLS regression accounts for the precision of the regional model and the effect of the record length on the variance of skewness estimators, GLS regression also considers the effect of cross-correlation of the skewness estimators on the sampling characteristics of the estimators. As it will be explained later, this is an important issue for the California regional skew study. The new Bayesian GLS regression procedures extend the GLS

regression framework by also providing a description of the precision of the estimated model error variance, a pseudo analysis of variance and enhanced diagnostic statistics. (See also Griffis and Stedinger, [2007].) A Bayesian GLS regional skew analysis was employed in the regional skew study recently completed to support flood frequency studies for the Southeastern United States [Veilleux, 2009; Weaver *et al.*, 2009; Feaster *et al.*, 2009; and Gotvald *et al.*, 2009].

Similar to the regional skew study performed in the Southeastern United States, the California regional skew study described here illustrates the use of the Bayesian GLS framework to support flood frequency analysis. However, the statistical procedures used in the Southeastern United States regional skew study were adapted and extended to address concerns that arose in the analysis of the California data set. Prior to performing the regional skewness analysis in California, a low outlier test (Expected Moments Algorithm) was employed with the California annual peak flow records and subsequently those records were adjusted, resulting in modified at-site skewness estimators. Also, the extremely large cross-correlations derived from the California annual peaks required special attention in the analysis. The extended Bayesian GLS regression framework employed in this study addresses the effects of both of these concerns. An extended Bayesian WLS-GLS analysis is then used to derive a regional skew model for California, which reflects California's unique hydrology, and to estimate its precision. Veilleux [2009] and Chapter 2 provide a discussion of the Bayesian GLS framework, while this chapter provides the details of the extended Bayesian WLS-GLS analysis in the following sections.

3.2 California Data

This section describes the California data, including the available basin characteristics. Also discussed in this section is a redundant site analysis, as well concerns with the Eastern

Sierra – Lahontan Desert sites. Finally, a model of cross-correlations of concurrent annual peaks is developed.

3.2.1 Overview of Data

This study is based upon annual peak flow data from 192 stream flow gauges (sites) located in California that were recommend by the United States Geological Survey (USGS), as well as the United States Army Corp of Engineers (USACE). The annual peak flow data is accessible on the USGS National Information System: Web Interface (NWISWeb). Each site is cataloged by a unique USGS eight or nine digit Hydrologic Unit Code (HUC), which is referred to in this study as a 'USGS site number' or simply just a 'site number'. In addition to the USGS site number, each site is also assigned a unique index number for this study ranging from 1 to 192. A list of the 192 sites can be found in Appendix A.

It is important to note that no sites were included from the southern Eastern Sierra region or the Lahontan desert region of California. There were 6 sites located along the eastern edge (or backside) of the Sierra as well as 9 sites scattered in the Lahontan Desert. By investigating a California annual average precipitation map [FRAP, 2000], as well as a California land cover map [FRAP, 2003], it is clear that this is a region distinctly different from the rest of the state. Thus, for the California regional skew study, this region will not be included, and consequently the results produced in the regional skew study will not apply to the southern Eastern Sierra and Lahontan Desert (ESLD) Region. Appendix B provides a list of 17 ESLD sites which were not included in the regional skew study (192 sites – 17 ESLD = 175 sites).

Site 11428000 is removed due to the poor fit of its annual peak flow record to the assumed underlying log-Pearson Type 3 (LP3) distribution. (Bulletin 17B recommends the LP3 distribution for flood flow frequency studies.) Below, Figure 3.1 shows the normalized

probability plot for Site 11428000. If the site fit the LP3 distribution well, the peak annual flows would form a straight line. As shown in Figure 3.1, the points do not resemble a straight line. Thus, the site was removed from the regional skew study (175 sites – 1 poor LP3 fitting site = 174 sites).

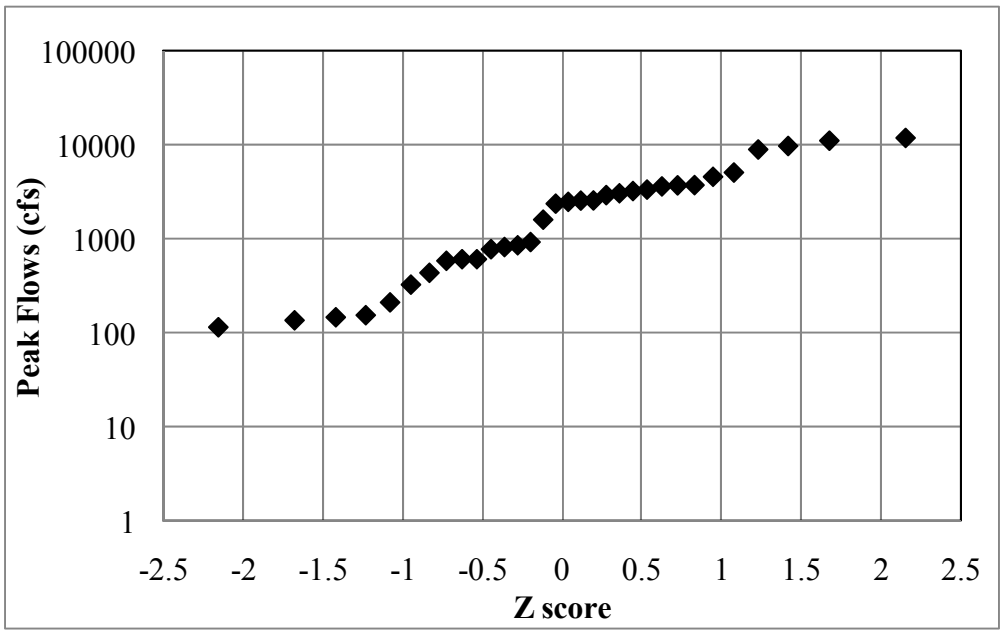


Figure 3.1: Probability plot for annual maximum peaks for USGS Site 11428000 with 32 years of record

In addition to the peak flow data, basin characteristics for the remaining 174 sites were provided by the USGS and the USACE. Table 3.1 lists the available basin characteristics.

Table 3.1: Basin characteristics for California annual maximum study

General Category	Basin Chacteristics
Location of Basin:	•Latitude of Centroid (decimal degrees) •Longitude of Centroid (decimal degrees) •Distance to the Coast, (miles)
Basin Area:	•Drainage Area, DA (square miles)
Basin Length:	•Basin Perimeter Length (miles)
Basin Slope:	•BSLDEM30M
Basin Elevation:	•Average Basin Elevation (feet) •Maximum Basin Elevation (feet) •Minimum Basin Elevation (feet) •Basin Outlet Elevation (feet) •Basin Relief Elevation (feet) •High Elevation Index, HELIND (%) (% of basin above threshold, where threshold = 3000, 4000, 5000, 6500)
Basin Precipitation:	•Average Annual Precipitation (inches)
Basin Temperature:	•Maximum January Temperature (°F) •Minimum January Temperature (°F)
Basin Coverage:	•Impervious Surface Coverage (%) •Forest Coverage (%) •Lake Coverage (%)
Physiographic Provinces:	•North Coast •Sierra •Central Valley •South Coast

The basin characteristics provided in Table 3.1 include percent of basin contained within physiographic provinces, as well as the more standard characteristics such as location of basin centroid, drainage area, main channel slope, and basin elevation.

3.2.2 Introduction to California Hydrology

The hydrology in California is extremely complex as there are large extremes in terrain and climate. In general, California does not experience much precipitation during the summer and fall months (June through October). However, in late November and early December, there is a shift in the jet stream over the Pacific Ocean which causes both rain and snow in California, depending on elevation. At the beginning of this period when the shift first occurs, the large storms do not necessary produce large peak flows, as there is a high infiltration capacity. It is not until after the infiltration capacity is exceeded that surface flow occurs. As a result, most of the large-scale flooding events in California are due to either large storms arriving one after the other or a large rain falling on soils previously saturated with snowmelt [Mount, 1995].

Mount [1995] explains that in central and northern California, for watersheds that are above 5500 *ft* most of the precipitation occurs as snow, and as such the snow packs act as water storage systems. Thus, during the winter season from mid-November to early April, the colder temperatures and increased precipitation allow for a thick snow pack. However, it is important to note that the snow pack growth does get disrupted by warm mid-Pacific storms which produce heavy rain and cause snowmelt. In early April, the temperatures begin to rise and the snowpack starts to melt, usually resulting in its complete melt by early July. Mount [1995] notes that it takes several months to melt a deep snow pack, so the spring (*i.e.* snowmelt) hydrographs are very different from rainfall hydrographs and their peak discharges tend to be much less.

3.2.3 Flood Peaks in California

The time of year of occurrence of flood peaks as a function of mean basin elevation was computed for the 158 California gauging stations and plotted in Figure 3.2. The *y*-axis holds

the average month of occurrence of annual peak discharge where 1 = October, 2 = November, 3 = December, … 6 = March, and so on. It is important to note that floods between August and September were ignored. For those sites with mean basin elevation less than 4,000 *ft*, the average months of occurrence of annual peak floods are in January and February (months 4 and 5, respectively). As mean basin elevation increases, so too does the average month of occurrence. When mean basin elevation is between 6,000 *ft* and 8,000 *ft*, the majority of the sites have their average month of occurrence between March and April (months 6 and 7, respectively). When mean basin elevation is greater than 8,000 *ft* the average months of occurrence of peak floods are between April and June (months 7 and 9, respectively). Figure 3.2 illustrates a fundamental hydrologic change in peak flows with mean basin elevation. However, by using average month of occurrence as a signal variable, the difference between a November flood and a February flood could be exaggerated. It is possible that the signal in peak floods could be due more to season than to a specific month.

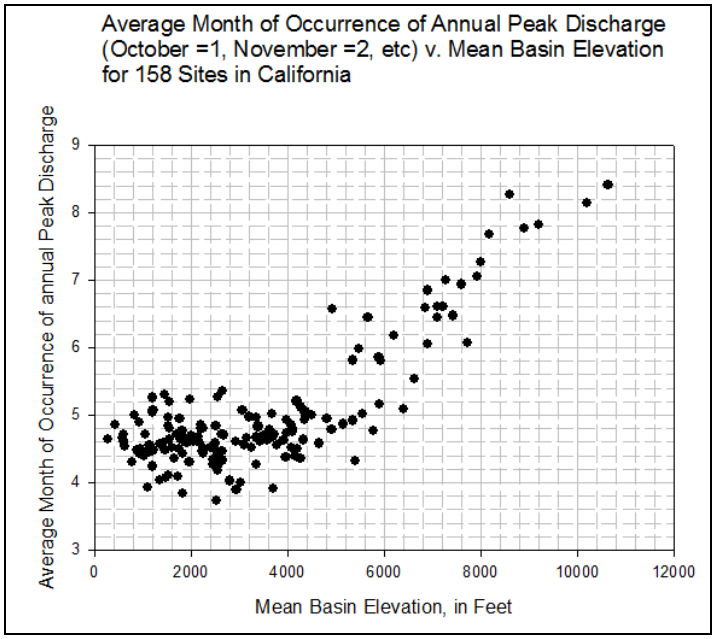


Figure 3.2: Average month of occurrence of annual peak discharge versus mean basin elevation for the 158 California gauge peaks used the California regional skew study. (Figure courtesy of Charles Parrett and Nancy Barth of the Sacramento, CA USGS WSC; Parrett *et al*., 2011)

Thus, Figure 3.3 shows the proportion of winter peak versus mean basin elevation, where a winter peak is one that occurs between November and March. This figure draws attention to the fact that sites with a mean basin elevation below 4000 *ft* have between 80% and 100% of their peak flows in the winter. This is percentage then decreases quickly as mean basin elevation increases. Sites with a mean basin elevation between 4000 *ft* and 8000*ft* have between 80% and 20% of their peak flows in the winter, and those sites with mean basin elevations above 8000 *ft* have 20% or less of their peak flows in the winter. Thus, similar to Figure 3.2, Figure 3.3 illustrates a fundamental hydrologic change in peak flows due to mean basin elevation. One drawback to using season of occurrence as a signal variable is that the difference between a March 30th peak and an April 1st peak is exaggerated. However, by considering Figure 3.2 and Figure 3.3 together, it is clear hydrology changes with mean basin elevation. In general, at both low and high elevation sites winter floods are rainfall driven. However at many high elevation sites spring floods are the results of snow melt events, or perhaps rain and snowmelt, while at the low elevation sites floods continue to be rainfall driven. It is also important to note that at sites with mean basin elevations greater than 8,000 *ft*, a mixed-population analysis may be needed to best explain the effect of the rain-snow interaction.

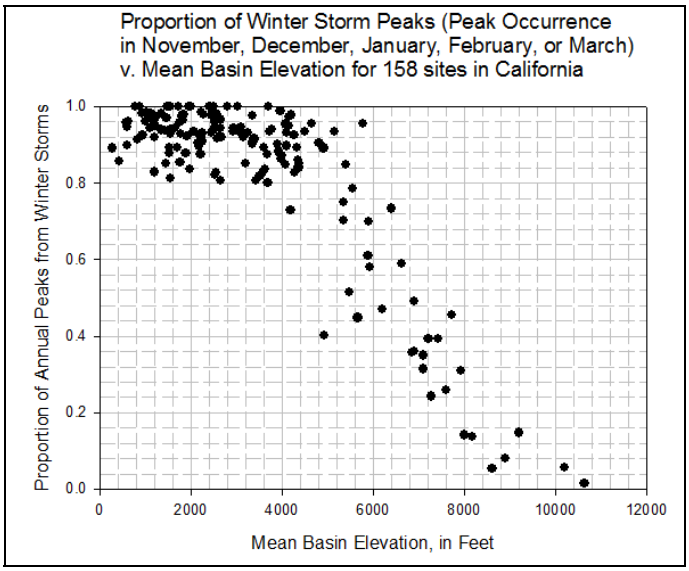


Figure 3.3: Proportion of annual peak discharge from winter storms versus mean basin elevation for the 158 California gauge peaks used the California regional skew study. (Figure courtesy of Charles Parrett and Nancy Barth of the Sacramento, CA USGS WSC; Parrett *et al*., 2011)

3.2.3.1 Understanding Floods at Low and High Elevation Sites in California

In order to understand better how the hydrology of California varies with respect to elevation, two sites with very different terrain are investigated. Site 11315000 is located in the Eldorado National Forest in the Sierra-Nevada mountain range east of Sacramento and south of Lake Tahoe. It has a mean basin elevation of 7414 *ft*, a drainage area of 21 mi^2, and is located 139 *mi* from the coast. Site 11159200 is located in the coastal hills just east of Santa Cruz, CA. It has a mean basin elevation of 1001 *ft*, a drainage area of 28 mi^2, and is located 6.2 *mi* from the coast. Figure 3.4 is a satellite image of California depicting the terrain and location of Site 11315000 and Site 11159200.

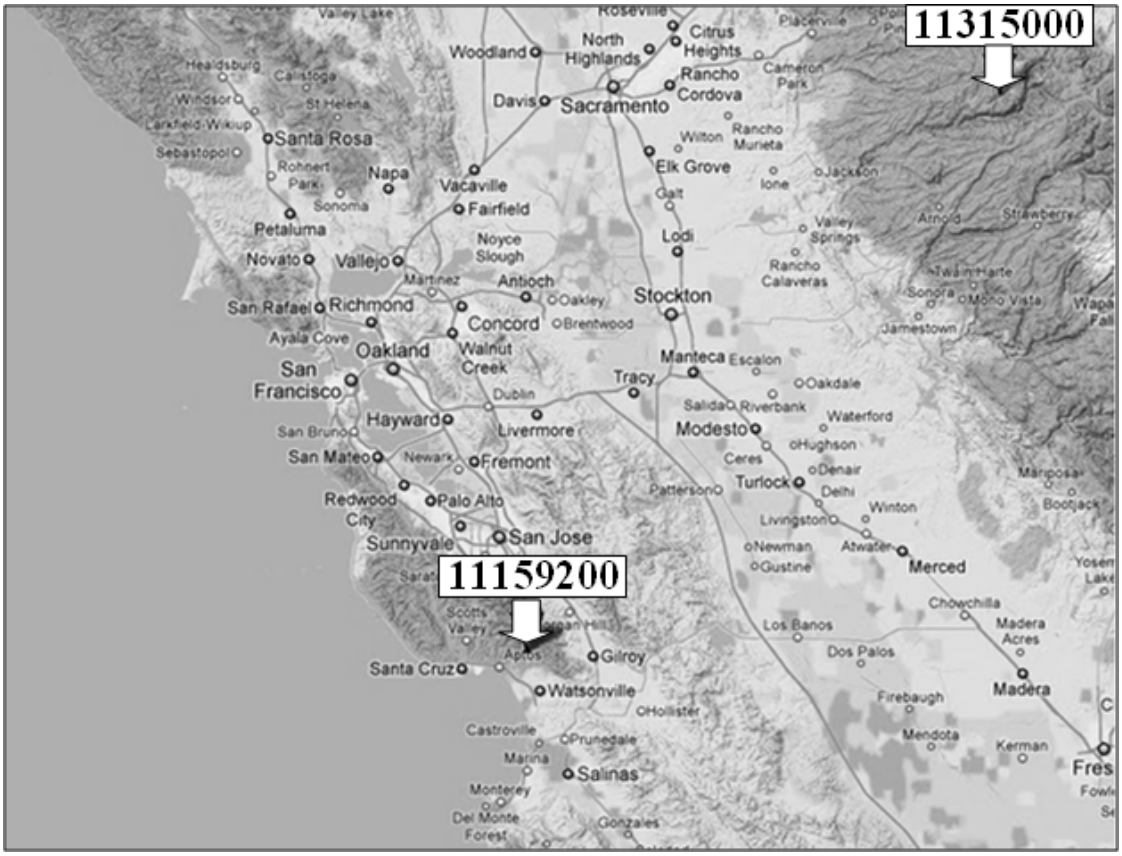
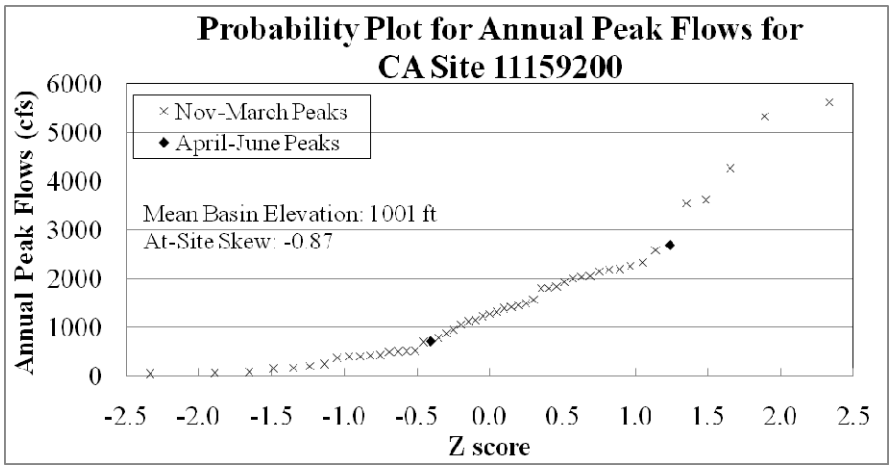


Figure 3.4: Satellite image of California depicting the location of the centroids of two basins from the CA regional skew study, Site 11315000 and Site 11159200 (Satellite Image Created From Google Maps).

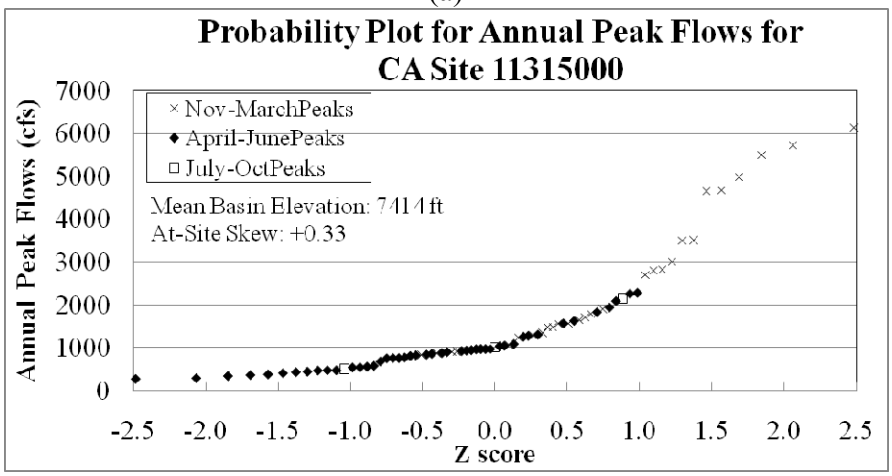
Figure 3.5 displays probability plots for the annual peak flow records for low elevation Site 11159200 (Figure 3.5a) and high elevation Site 11315000 (Figure 3.5b). At low elevation sites in California, the mean annual peaks are almost all due to rainfall events, which have a negative log-space skews (Site 11159200 has an at-site log-space skew of -0.87). Rainfall events in California in areas with Mediterranean climate have a thin tail. The larger storms are not that much larger than more frequent events. This can be seen in Figure 3.5a in which all but two of the annual peaks occur in the winter (*i.e.* rainy) season between Nov-March. Also, Site 11159200 has very thin tails at the small and large ends, as most of the peaks are grouped between 1000 *cfs* and 2500 *cfs*. On the other hand, at high elevation sites, the majority of the mean annual peaks

are due to snow melt. This can be seen in Figure 3.5b, as the majority of the annual peaks occur in the spring snow melt months between April and June. However, on occasion when a large rain event occurs it can produce a large annual maximum. This can be seen in Figure 3.5b, as the largest twelve peaks are all winter rainfall events. This mixture of events results in a positive at-site log-space skew at Site 11315000 of +0.33.

This explains the trend in skew from below 4,000 ft where the annual floods are almost all rainfall events, to high elevations where the annual floods are mostly snowmelt with a few large rainfall events. The rainfall distribution at high elevation sites is also relatively tame, but does have a larger standard deviation that the corresponding snowmelt peaks. Thus, when the two distributions, snowfall and rainfall, are combined, the combination has a positive log-space skew. However, a 3-parameter log Pearson type 3 (LP3) distribution with a positive skew is likely to significantly exaggerate the risk of very extreme flood flows at these higher elevation sites.



(a)

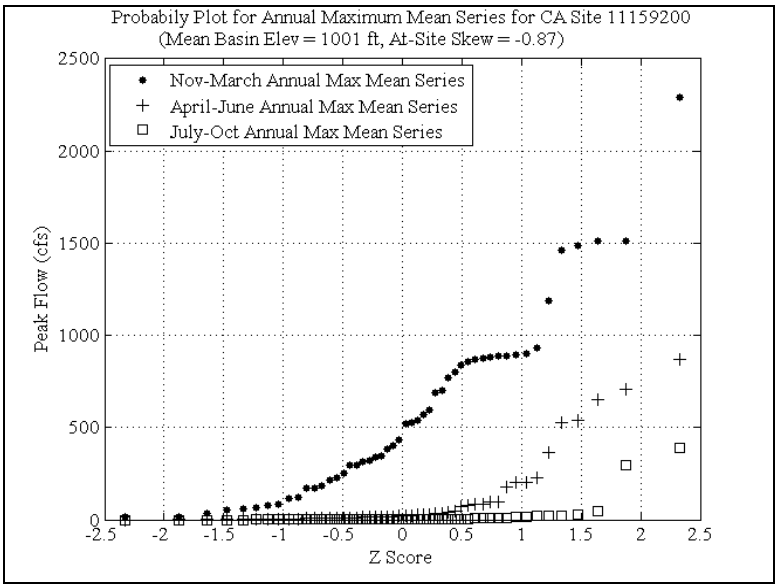


(b)

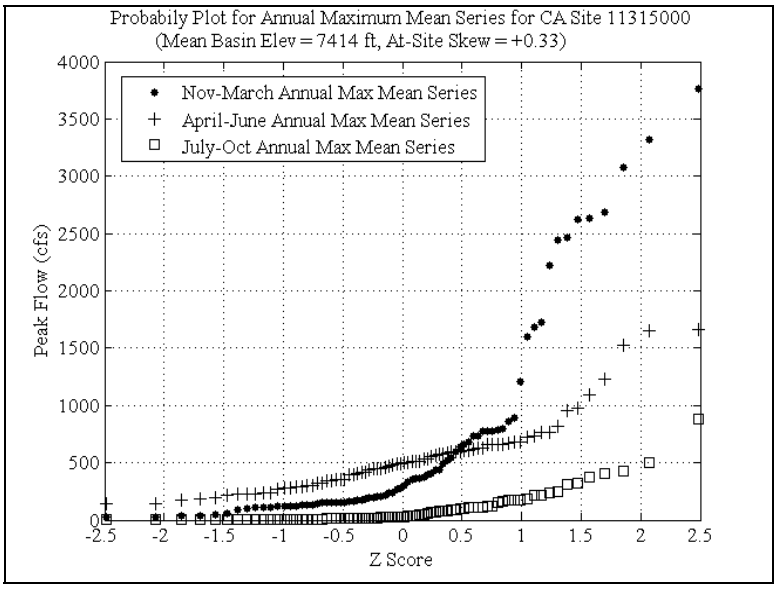
Figure 3.5: Normal probability plots for the annual peak flows for two sites in the California regional skew study. The annual peak flows from Nov and March are displayed as x's, the peaks between April and June are displayed as dots, and the peaks between July and October are displayed as squares. The graph on top (a) is the probability plot for CA site 11159200, which has an elevation of 1001 *ft* and an at-site log-space skew of -0.87. The graph on the bottom (b) is the probability plot for CA site 11315000, which has an elevation of 7414 *ft* and an at-site log-space skew of +0.33.

Figure 3.6 contains the normal probability plots for the seasonal annual maximum mean daily series for the two CA sites in Figure 3.4. Each of the three series in the graphs in Figure 3.6 represent a different season where winter spans November through March, spring spans April through June, and summer spans July through October. In order to generate these plots, the daily mean flow time-series for each day in the period of record was downloaded from the USGS

National Water Information Service: Web Interface (USGS NWISWeb). From these time-series, the maximum daily mean for each month is determined. Finally for each season, the seasonal max is determined from the monthly maximum daily mean. Thus, each of the three series represents the maximum daily mean in each season in each year. Figure 3.6a is the probability plot for low elevation site 11159200, which has a mean basin elevation of 1001 *ft* and an at-site log-space skew of -0.87. Figure 3.6b is the probability plot for high elevation site 11315000, which has a mean basin elevation of 7414 *ft* and an at-site log-space skew of +0.33. At both the low elevation coastal site and the high elevation mountain site, the largest annual maximum mean daily peaks occur in the winter season followed by the spring season and then the summer season.



(a)



(b)

Figure 3.6: Normal probability plots for two sites in the California regional skew study created from the mean daily flows at each of the sites. The annual maximum mean daily peaks between Nov and March are displayed as dots, the peaks between April and June are displayed as plus signs, and the peaks between July and October are displayed as squares. The top graph (a) is the probability plot for CA site 11159200, which has an elevation of 1001 *ft* and an at-site skew of -0.87. The bottom graph (b) is the probability plot for CA site 11315000, which has an elevation of 7414 *ft* and an at-site skew of +0.33.

3.2.4 California Hydrology Conclusions

The hydrology of California indicates that the annual maximum flood flows in California are realted to the mean basin elevations. It is shown that gauge sites with mean basin elevation below 4,000 *ft* have their maximum annual floods driven by a different hydrologic mechanism than those basins with elevations above 4,000 *ft*. Those sites with mean basin elevations below 4,000 *ft* have their maximum annual floods driven by rainfall events. However as the mean elevation of basins increases above 4,000 *ft,* the interaction of rainfall and snowmelt events increasingly effects the maximum annual floods.

3.3 California Data Analysis

This section discusses the data analysis performed before a regional skew regression model is formed. First, a redundant site analysis is conducted and then a model of the cross-correlations of concurrent annual peaks is developed.

3.3.1 Redundant Sites

In the Southeastern United States regional study [Gruber and Stedinger, 2008; Veilleux, 2009; Feaster *et al.,* 2009; Gotivald *et al.,* 2009; Veilleux, 2009; Weaver *et al*., 2009], it was discovered than many pairs of gauges were measuring essentially the same flood series because the larger basin entirely contained the smaller basin. This is referred to here as redundancy. That study uses metrics to identify redundancy that were developed in Gruber and Stedinger [2008]. A redundancy check was performed on the California data set to see if there were similar issues.

The normalized distance *ND* between two sites *i* and *j* is defined to be

$$ND = \frac{D_{ij}}{\left[A_i A_j\right]^{1/4}} \tag{3.1}$$

where D_{ij} is the distance between centroids of basin i and basin j, and A_i and A_j are the drainage areas for basin i and basin j.

The fourth root is required to make *ND* dimensionless. The drainage area ratio *DAR* is defined as

$$DAR = Max\left[\frac{A_i}{A_j}, \frac{A_j}{A_i}\right] \tag{3.2}$$

Simple examples suggest those station-pairs with *ND* less than 0.5 are likely to be physically nested [Veilleux, 2009]. Moreover, if in addition *DAR* is less than 5, then the nested sites are most likely redundant. If *DAR* is very large, then even if two sites are nested, they are unlikely to be redundant. This is due to the fact that one basin's drainage area is so much larger than the other that they may not respond to storms in the same way. Observed cross-correlations of annual peak flows were consistent with this hypothesis in the Southeastern U.S. study.

The 174 sites for use in the California regional skew study were screened for redundant sites using *ND* and *DAR*. Site-pairs are considered to be redundant if $ND < 0.5$ and $DAR < 5$. This resulted in only 14 site-pairs that appeared to be redundant; as a result of an examination of those site pairs, 14 sites are suggested to be removed from the study, one from each pair. Also, the two largest Army Corp of Engineers (USACE), Sacramento River at Keswick and Feather River at Oroville, as measured by drainage area, are removed as they contained many other drainage areas in the study and were used to extend the records of other USACE sites. These 16 sites (see Appendix C) are removed from the regional skew study (174 sites – 16 redundant sites = 158 sites).

3.3.2 Cross-Correlation Model of Concurrent Annual Peaks

In order to implement the Reis *et al.* [2005] Bayesian GLS regression framework (See Chapter 2), the correlations among skewness estimators must be described. This is a part of the GLS regression framework, in which the relationships between estimators for different sites are measured by their cross-correlations. As the true skew values are not know, these correlations can be approximated as a function of the cross-correlation of the peak flood flows as developed in Martins and Stedinger [2002].

A cross-correlation model for the annual maximum flows in California to be used in conjunction with the relationship derived by Martins and Stedinger [2002] for the cross-correlation of the skew coefficient estimators is developed using 21 sites with greater than 65 years of concurrent record and no censored peaks (Appendix D provides a list of 21 sites used in cross-correlation model). None of the key dam sites identified by the USCOE are used in this analysis because annual peak discharges at those sites are estimated. Various models relating the cross-correlation of the concurrent annual peak discharge at two sites, ρ_{ij}, to various basin characteristics were considered. In general, a logit model, termed the Fisher Z Transformation, Z = log[(1+*r*)/(1-*r*)] provides a convenient transformation of the sample correlations r_{ij} from the (-1, +1) range to the $(-\infty,+\infty)$ range. Table 3.2 contains a summary of the two best cross-correlation models, Model A the constant model and Model B which depends on the distance between basin centroids.

Table 3.2: Summary of cross-correlation regressions for California annual peak flow regional skew study (n = 21 sites, 159 site-pairs). σ_δ^2 is the model error variance, R_δ^2 is the Pseudo R_δ^2, and ERL is the effective record length.

Model	Beta Parameters b_0	b_1	b_2	σ_δ^2	R_δ^2	ERL
A: $Z_{ij} = b_0$	0.69 (0.028)			0.11	0%	12
B: $Z_{ij} = \exp\left(b_1 - b_2 * D_{ij}\right)$		0.27 (0.030)	-0.0037 (2.1E-04)	0.02	81%	52

Based on the results provided in Table 3.2, Model B is adopted for the estimation of the cross–correlations of concurrent annual peak discharge at two stations. The ordinary least squares analysis indicates that Model B is as accurate as having 52 years of concurrent annual peaks from which to calculate a cross-correlation, which is much improved over the 12 years of concurrent annual peaks achieved with the Constant Model. Model B, which uses the distance between basin centroids as the only explanatory variable, is

$$\rho_{ij} = \frac{\exp\left(2Z_{ij}\right) - 1}{\exp\left(2Z_{ij}\right) + 1} \tag{3.3}$$

where

$$Z_{ij} = \exp\left(0.27 - 0.0037 * D_{ij}\right) \tag{3.4}$$

Figure 3.2 shows the fitted relation between the Fisher Z transformed cross-correlations and the distance between basin centroids together with the plotted sample data from the 159 station pairs of data. Figure 3.3 shows the functional relation between the un-transformed cross-correlation and distance between basin centroids in both California and the Southeastern United States.

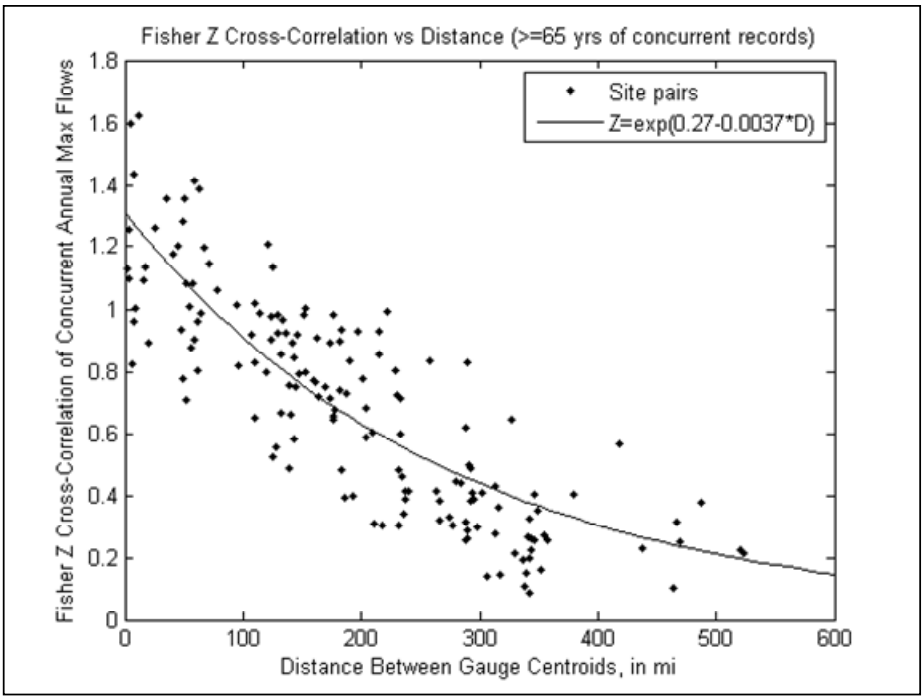


Figure 3.7: Relationship between Fisher transformed cross-correlation (Z) of the logs of annual peak discharge and distance between basin centroids for 159 station pairs in California with at least 65 years of concurrent annual peak flows.

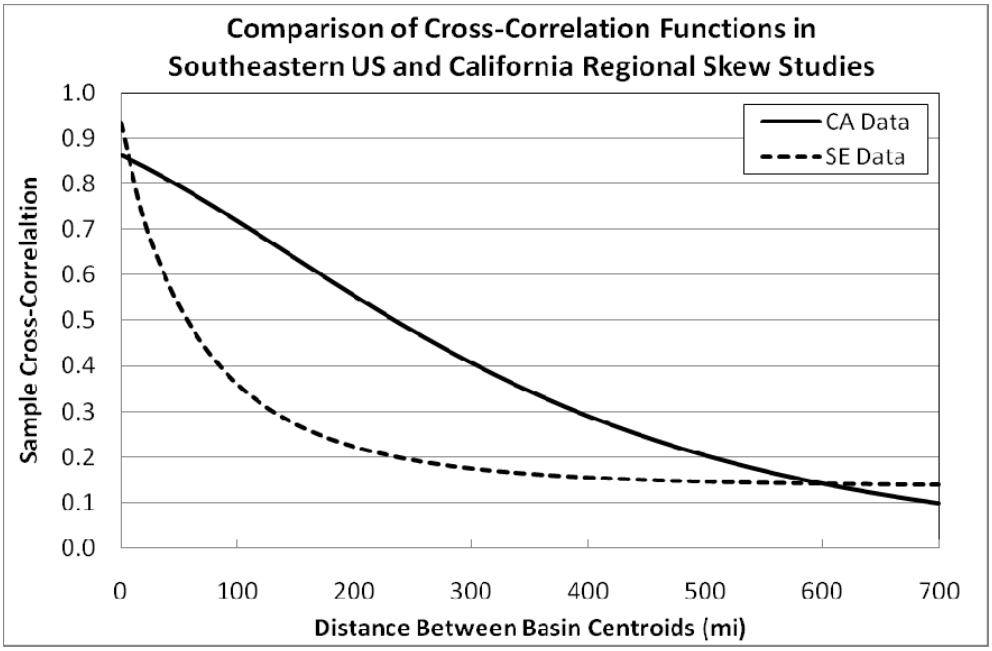


Figure 3.8: Relation between un-transformed cross-correlation of logs of annual peak discharge and distance between basin centroids based on data from 159 station pairs in California and 1317 station pairs in the Southeastern United States.

As shown in Figure 3.8, the distance between basin centroids yields more gradual decreases in the cross-correlation between annual peak flows in California than in the Southeastern United States. Thus, in general the cross-correlations in California are much larger than those in the Southeastern U.S. This mostly reflects the character of the storms that cause large floods in the two regions.

In both California and the Southeastern U.S. regional skew studies, the cross-correlation model is used to estimate the site-to-site cross-correlations for concurrent annual floods at all pairs of sites in the studies. Figure 3.5 is a histogram of the relative frequencies of the estimated cross-correlations among the 158 sites in the California data set and the 342 sites in the Southeastern U.S. data set.

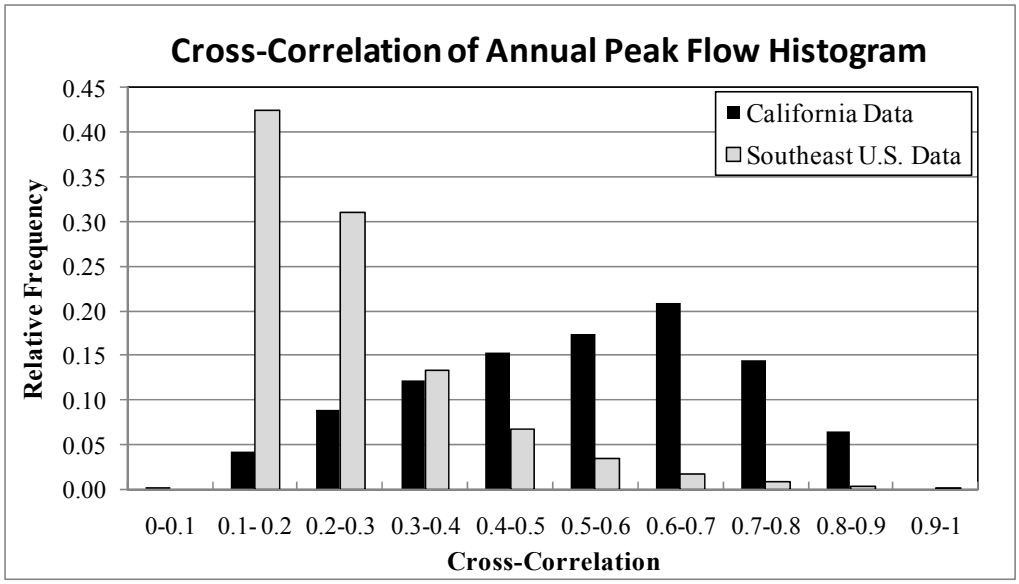


Figure 3.9: Histogram showing relative frequency of calculated cross-correlation values for both the California data set (158 sites = 12,403 site-pairs) and the Southeastern U.S. data set (342 sites = 58,311 site-pairs).

Figure 3.9 clearly shows that the drastic difference in cross-correlation models shown in Figure 3.8, results in drastically different cross-correlations between site-pairs. For the

Southeastern U.S. the largest frequencies of cross-correlations occur between 0.1-0.2 and 0.2-0.3. Alternatively, the largest frequencies of cross-correlations for the California data set occur between 0.6-0.7 and 0.5-0.6. The effects of these large cross-correlations will be discussed in Section 3.3.

Below, Figure 3.10 is a histogram of the relative frequencies of the distance between basin centroids among the 158 sites in the California data set and the 342 sites in the Southeastern U.S. data set.

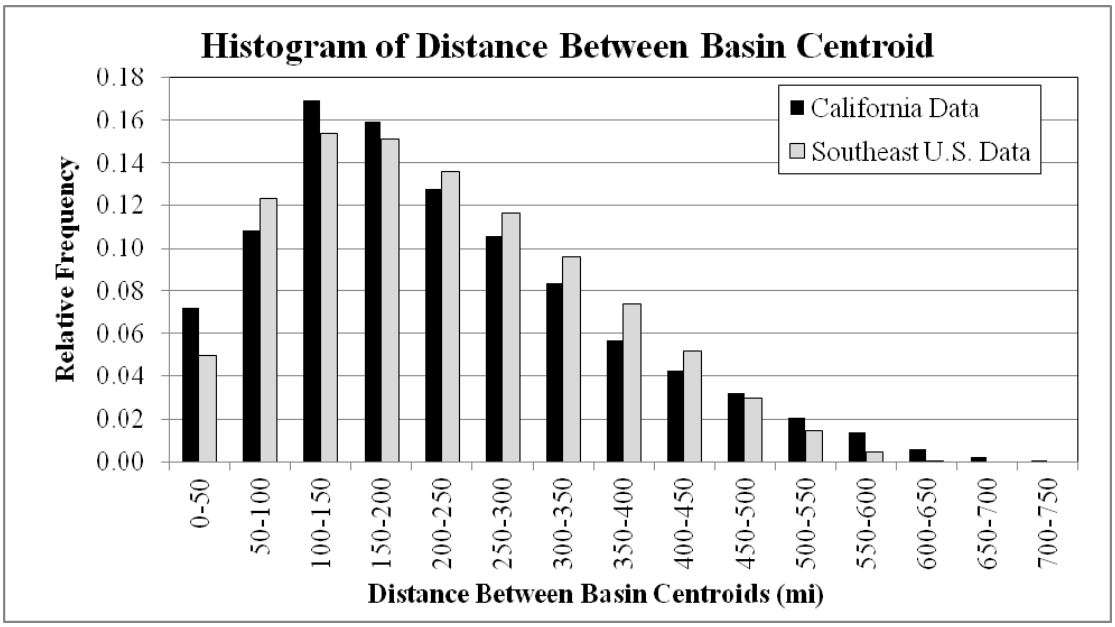


Figure 3.10: Histogram showing relative frequency of distance between basin centroids for both the California data set (158 sites = 12,403 site-pairs) and the Southeastern U.S. data set (342 sites = 58,311 site-pairs).

As shown in Figure 3.10, the California data has more basin pairs located between 0 and 200 miles, while the Southeastern data set has more basin pairs located between 200 and 450 miles. These distances are then compared to the cross-correlation as a function of distance for each study as shown in Figure 3.8. These two figures support the results in cross-correlation

histogram in Figure 3.9. The California data set has many of its gauge sites located close together and those close distances have the largest cross-correlations, thus causing there to be a large frequency of high cross-correlations in California. Similarly, the gauge sites in the Southeast tend to be located at greater distances apart, and have smaller cross-correlations at those distances, thus resulting in smaller frequency of high cross-correlations.

3.4 Extended Bayesian GLS Regional Skewness Methodology

The Southeastern United States regional skew analysis illustrates how a Bayesian Generalized Least Squares (B-GLS) analysis would typically proceed [Veilleux, 2009; Weaver and others, 2009; Feaster and others, 2009; and Gotvald and others, 2009]. Compared to the Southeastern United States, the cross-correlations between annual peak discharges in California are often large, as shown in Figure 3.8 and Figure 3.9. When a B-GLS analysis is attempted with the California data set, consistent results are not obtained because of these high cross-correlations. A Bayesian GLS analysis seeks to exploit the cross-correlations among the sample skews to obtain the best estimates of model parameters possible. If the cross-correlations are large, the GLS estimators can become relatively complicated as a result of the effort to find the most efficient estimator of the parameters. The model (Equation 3.4) developed to generate the cross-correlations of concurrent annual peaks in California is equivalent to an at-site record with 53 years of data and describes the overall structure of the California data set. But, the precision of the cross-correlation estimates between any two particular sites is not of sufficient precision to justify the sophisticated weights (both positive and negative) that the Bayesian GLS analysis generates. Section 3.3.1 demonstrates when a Bayesian-GLS regression falls apart based on different cross-correlation models.

3.4.1 Failure of Bayesian GLS

This section uses data based on the California annual maximum peak flow data set to illustrate when Bayesian-GLS regression for estimating a regional skew model falls apart. In order to illustrate when B-GLS regression falls apart, four cross-correlations models are applied to the California annual maximum peak flow data set. These four cross-correlation models are presented in Table 3.3.

Table 3.3: Cross-Correlation models for California annual maximum peak flow data set for B-GLS failure test.

Cross-Correlation Regression Parameters for Model: $\rho_{ij} = \frac{\exp\left(2Z_{ij}\right)-1}{\exp\left(2Z_{ij}\right)+1}$ where $Z_{ij} = \exp\left(b_1 + b_2 D_{ij}\right)$		
Model	**b_1**	**b_2**
Original	0.27	-0.0037
Model 2	0.27	-0.0090
Model 3	0.27	-0.0300
Model 4	0.27	-0.5000

The Original Model is the best fit cross-correlation model for the California data set and is given in Equation 3.4 in Section 3.4.2. The other three models in Table 3.3 all have the same form as the Original Model, however the regression coefficient in front of the distance term has been increased (in absolute value). As shown in Figure 3.11, increasing the value of that regression coefficient causes the slope of the cross-correlation model to decrease more quickly, thus reducing the cross-correlations.

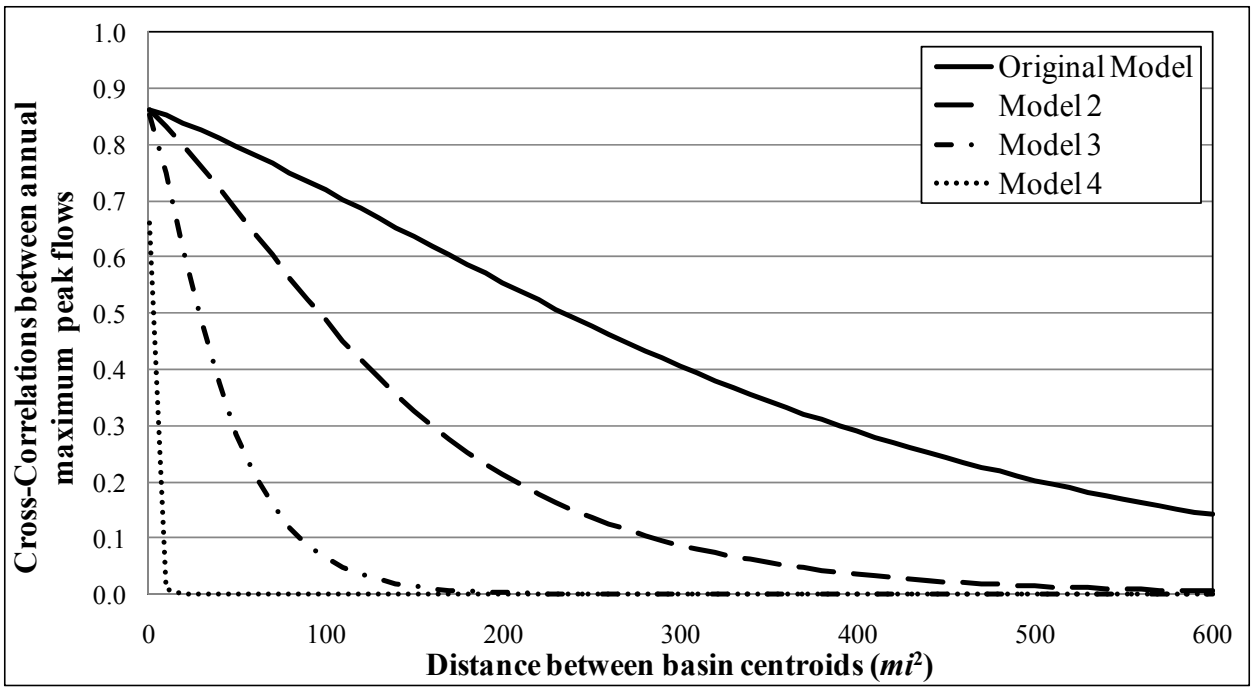


Figure 3.11: Cross-Correlation models for California annual maximum peak flow data set for B-GLS failure test. The four models in this figure correspond to the four models described in Table 3.3.

As shown in Figure 3.11, Model 2, Model 3, and Model 4 substantially reduce the cross-correlation as a function of increasing distance. Table 3.4 provides summary statistics for the cross-correlations of the California data set when each of the cross-correlation models in Table 3.3 (and Figure 3.11) is applied to the data set. These statistics are calculated for all 158 sites in the California annual maximum data set. Table 3.4 also provides summary statistics for the cross-correlations of the Southeastern U.S. data set for the cross-correlation model used for the Southeastern regional skew model. The cross-correlations statistics for the Southeastern U.S. serve as a comparison for the cross-correlations for the California data set, as the B-GLS methodology was successfully applied to develop a regional skew model for the Southeast U.S.

Table 3.4: Cross-correlation summary statistics for all 158 sites in the California annual maximum data set for all cross-correlation models provided in Table 3.3. Also included are the cross-correlation summary statistics for the 342 sites in the Southeast annual maximum data set for the cross-correlation model used in the Southeastern regional skew model.

Model	Average	Std. Dev.	Median	Mode	Max	Min	% of Original Average ρ
Original	0.54	0.18	0.56	0.08	0.86	0.08	
Model 2	0.27	0.22	0.22	0.00	0.86	0.00	51%
Model 3	0.06	0.13	0.00	0.00	0.85	0.00	11%
Model 4	0.01	0.05	0.00	0.00	0.81	0.00	2%
SE Original	0.26	0.13	0.22	0.14	0.91	0.14	N/A

Table 3.4 shows the large differences between the Original Model and the other three models. As shown in Figure 3.11 and confirmed in Table 3.4, the average cross-correlation in the California data set is significantly lower for Models 2 through 4. The average cross-correlation in Model 2 is about half of the average cross-correlation the Original Model. Table 3.4 also shows the large difference between the Original Model for California and Original Model the Southeastern U.S. The average cross-correlation in California is over twice the average cross-correlation in the Southeastern U.S.

The cross-correlation models in Table 3.3 are then used to develop regional skew models. Table 3.5 contains the results for OLS, Bayesian-WLS (B-WLS), and B-WLS analyses for a constant regional skew model for California. The OLS and B-WLS models do not consider cross-correlation between annual peak flows. Table 3.5 presents four B-GLS models, one for each of the cross-correlation models in Table 3.3.

Table 3.5: Regional skew regression results for the constant model of California annual maximum flood data set. Standard deviations are presented in ( ), σ_δ^2 is the model error variance, ASEV is the average sampling error variance, AVPnew is the average variance of prediction for a new site.

Cross-Correlation Model	Regression Constant	σ_δ^2	ASEV	AVP_{new}
OLS Regression				
N/A	-0.31	0.24	0.001	0.23
	(0.04)	-		
B-WLS Regression				
N/A	-0.30	0.11	0.002	0.11
	(0.04)	(0.03)		
B-GLS Regression				
Original Model	-0.21	0.06	0.016	0.08
	(0.13)	(0.02)		
Model 2	-0.26	0.04	0.007	0.05
	(0.08)	(0.01)		
Model 3	-0.32	0.04	0.002	0.04
	(0.05)	(0.02)		
Model 4	-0.30	0.10	0.002	0.10
	(0.04)	(0.03)		

As shown in Table 3.5, considering the cross-correlations between annual maximum floods does impact the results of the regression analysis. The B-GLS model using the Original cross-correlation model results in the smallest (in magnitude) regression constant. By reducing the average cross-correlation, the regression constant increases (in magnitude). Model 4, which has an average cross-correlation of 0.01, as expected, very closely resembles the results from the OLS and B-WLS analyses which do not take into cross-correlation. Now the weights produced to perform each by each type of regression (OLS, B-WLS, B-GLS) can be compared.

The OLS weights are calculated as

$$\mathbf{W}_{OLS} = \left[\left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \right]^T \tag{3.5}$$

where $\mathbf{W}_{\text{OLS}}$ is an (n x k) matrix of OLS weights, $\mathbf{X}$ is an (n x k) matrix of basin characteristics with a first column of ones, n is the number of gauge sites, and k is the number of regression parameters. The WLS weights are calculated as

$$\mathbf{W}_{WLS} = \left[\left(\mathbf{X}^T \boldsymbol{\Lambda}_{WLS}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{WLS}^{-1} \right]^T \tag{3.6}$$

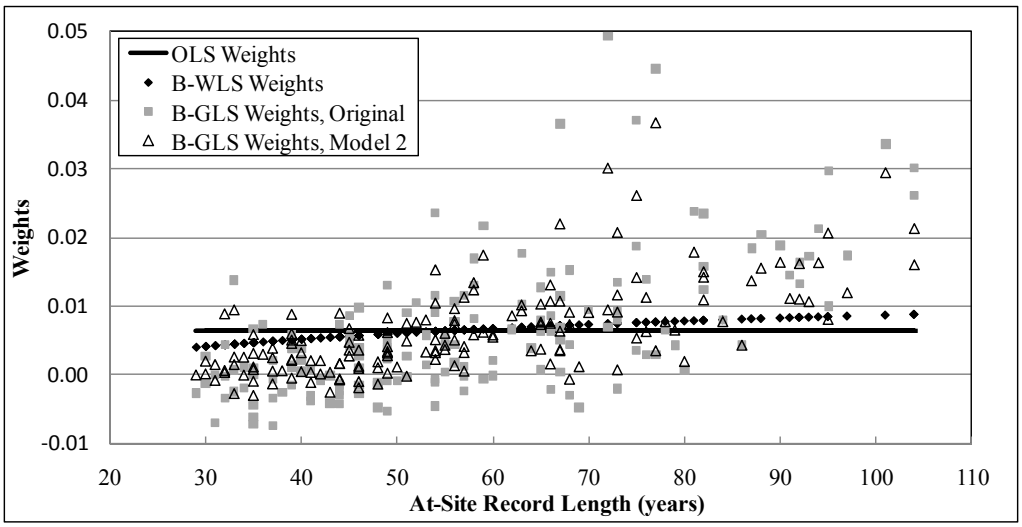
where $\mathbf{W}_{\text{WLS}}$ is an (n x k) matrix of WLS weights and $\boldsymbol{\Lambda}_{WLS}$ is an (n x n) WLS covariance matrix with the at-site variance on the diagonal and off diagonal components equal to zero. The GLS weights are calculated as

$$\mathbf{W}_{GLS} = \left[\left(\mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{GLS}^{-1} \right]^T \tag{3.7}$$

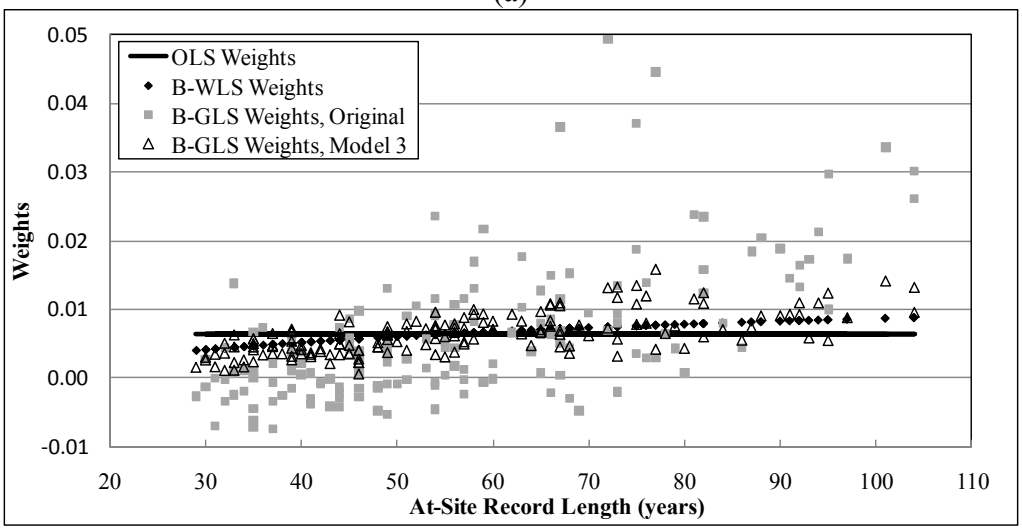
where $\mathbf{W}_{\text{WLS}}$ is an (n x k) matrix of GLS weights and $\boldsymbol{\Lambda}_{WLS}$ is an (n x n) GLS covariance matrix.

Figure 3.8 contains three graphs which compare the weights assigned to each gauge site in the regional skew regression by the OLS, B-WLS and B-GLS models. All three of the graphs contain the weights from the OLS, B-WLS and B-GLS when the Original cross-correlation model is considered. Figure 3.8a also includes the B-GLS weights when the Model 2 cross-correlation model is considered. Figure 3.8b includes the B-GLS weights when the Model 3 cross-correlation model is considered and Figure 3.8c includes the B-GLS weights when the Model 4 cross-correlation model is considered. The y-axis in Figure 3.8 contains the weights, while the x-axis is sorted by the number of years of at-site for each of the 158 gauge sites in the California data set.

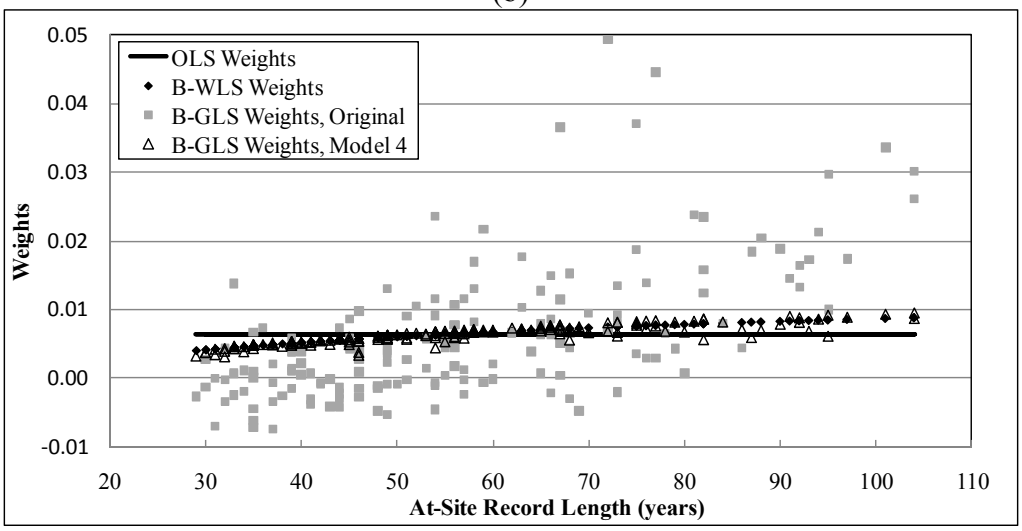
Figure 3.12: Comparison of weights from OLS, B-WLS, and B-GLS analyses for constant regional skew model for California annual maximum flood data set.


OLS Weights
B-WLS Weights
B-GLS Weights, Original
B-GLS Weights, Model 2
Weights
At-Site Record Length (years)

(a)


OLS Weights
B-WLS Weights
B-GLS Weights, Original
B-GLS Weights, Model 3
Weights
At-Site Record Length (years)

(b)


OLS Weights
B-WLS Weights
B-GLS Weights, Original
B-GLS Weights, Model 4
Weights
At-Site Record Length (years)

(c)

As shown in Figure 3.12 there are large differences between the B-GLS weights depending on which cross-correlation model is applied. As expected the OLS weights are constant for all gauge sites. This is due to the fact that OLS does not take into account at-site record length or cross-correlation, so each site in the study is weighted equally. The B-WLS analysis does not take into account cross-correlation, but it does factor in the at-site record length for each site in the study. Thus, B-WLS assigns large weights to those sites which with longer records. This accounts for the smooth increasing trend in weights as the at-site record length increases. The most complicated set of weights are the B-GLS weights, which account for record length and cross-correlation. While it is expected that in general those sites with the longest at-site record lengths will still have the largest weights, it is also expected that the B-GLS weights will not follow a smooth curve. It is expected that the cross-correlation of annual peak flows will introduce noise into the weights. However, as shown in Figure 3.12, there is a large amount of noise in the B-GLS weights when the Original cross-correlation model is used. Not only is there a lot of noise, but it is important to note that many of the weights are highly negative. By comparing Figures 3.12a, 3.12b and 3.12c, it is shown that as the average cross-correlation in the annual peak flows is decreased, the noise in the B-GLS weights decreases and the number of negative weights also decreases. When the average cross-correlation in the annual peaks flows approaches zero (Model 4), the B-GLS weights approximate the B-WLS weights. The negative weights that arise when using the Original Model to model the cross-correlations cause the B-GLS methodology to fail when developing regional skew models.

Figure 3.13 compares the OLS, B-WLS and B-GLS weights assigned to each gauge site in the Southeastern U.S. for the constant regional skew regression. By comparing Figure 3.13 to Figure 3.12, it is evident that the B-GLS weights in the Southeast U.S. are much less noisy the B-

GLS weights in California. The range of the B-GLS weights for the California Original Model is over five times larger than the range of the B-GLS weights for the Southeast U.S. Original Model. Also, 28% of the sites in the California data set have negative B-GLS weights for the Original Model, while only 2% of the sites in the Southeast U.S. data set have negative B-GLS weights.

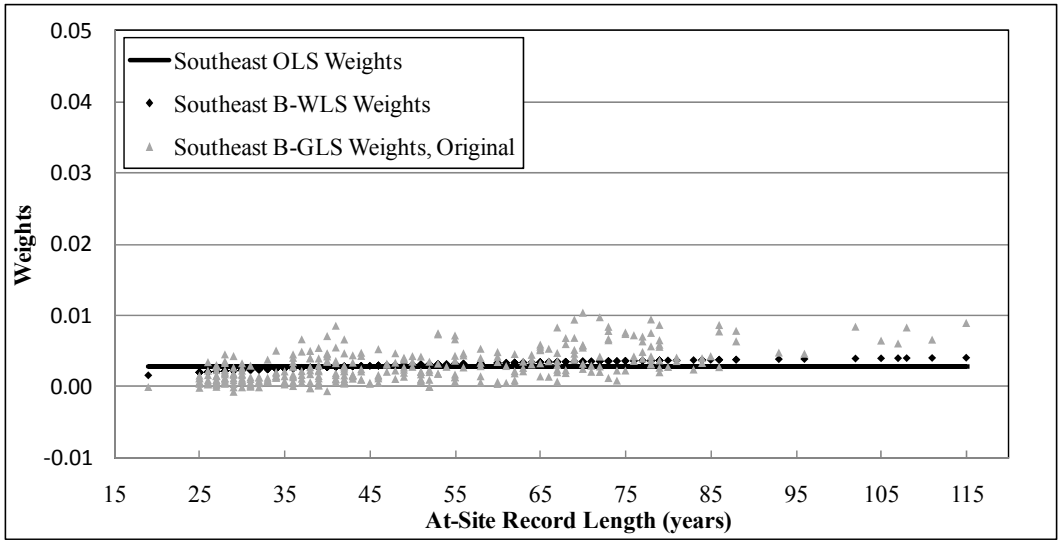


Figure 3.13: Comparison of weights from OLS, B-WLS, and B-GLS analyses for the constant regional skew model for the Southeastern annual maximum flood data.

In order to classify when B-GLS fails, a metric is developed which calculates the loss of efficiency when using B-GLS instead of B-WLS. This metric, Loss of Efficiency (LE), is calculated as

$$LE = \left(AVP_{WLS} - AVP_{GLS}\right)/AVP_{GLS} \tag{3.8}$$

where AVP_{WLS} is the average variance of prediction when considering a WLS regression and AVP_{GLS} is the average variance of predicition when considering a GLS regression. AVP_{WLS} is calculated as

$$AVP_{WLS} = \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n} \mathbf{x}_i^T * Var\left[\boldsymbol{\beta}_{WLS}\right] * \mathbf{x}_i\right) \tag{3.9}$$

where $\mathbf{x}_i$ is an (1 x k) vector of basin characteristics for site i, and $Var\left[\boldsymbol{\beta}_{WLS}\right]$ is the variance of the estimated regression parameters given the WLS weights and the GLS covariance matrix $\boldsymbol{\Lambda}_{GLS}$. Thus, the WLS weights are used, but it is assumed that cross-correlation between annual peak floods does exist. Thus, $Var\left[\boldsymbol{\beta}_{WLS}\right]$ is calculated as

$$Var\left[\boldsymbol{\beta}_{WLS}\right] = \mathbf{W}_{WLS}^{T} * \boldsymbol{\Lambda}_{GLS} * \mathbf{W}_{WLS} \tag{3.10}$$

Similarly, the AVP_{GLS} is calculated as

$$\text{AVP}_{\text{GLS}} = \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n} \mathbf{x}_i^{T} * Var\left[\boldsymbol{\beta}_{GLS}\right] * \mathbf{x}_i\right) \tag{3.11}$$

where

$$Var\left[\boldsymbol{\beta}_{GLS}\right] = \left(\mathbf{X}^{T} * \boldsymbol{\Lambda}_{GLS}^{-1} * \mathbf{X}\right)^{-1} \tag{3.12}$$

Another metric that can be used to determine when B-GLS fails is the root mean squared deviation (RMSD) between the B-WLS weights and the B-GLS weights. This metric measures the difference between the B-WLS weights and the B-GLS weights. It is standardized by the constant OLS weights.

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}\left(\mathbf{W}_{GLS} - \mathbf{W}_{WLS}\right)^{2}}{n\mathbf{W}_{OLS}^{2}}} \tag{3.13}$$

When the B-GLS weights are very different from the B-WLS weights the RMSD is large. If the B-WLS weights are equal to the B-GLS weights then the RMSD = 0. Table 3.6 contains both the LE and the RMSD metrics for the B-GLS regional skew regression for California based on cross-correlation models in Table 3.3. As a comparison, the loss of efficiency and RMSD metrics are also presented for the Southeast U.S. B-GLS regional skew regression.

Table 3.6: Loss of efficiency and RMSD metrics for B-GLS regional skew regression for California based on cross-correlation models in Table 3.3. As a comparison, the loss of efficiency and RMSD metrics are also presented for the Southeast U.S. B-GLS regional skew regression.

Model	Avg Cross-Corr	RMSD	LE
Original	0.54	1.45	36%
Model 2	0.27	0.94	24%
Model 3	0.06	0.37	10%
Model 4	0.01	0.10	1%
SE Original	0.26	0.67	18%

As shown in Table 3.6, the largest RMSD value and the largest LE percentage occurs when the Original cross-correlation model is used in the California B-GLS regression. As the average cross-correlation is decreased in Models 2 through 4, the RMSD and LE values also decrease. This is to be expected as the average cross-correlation in Model 4 is almost 0 and thus closely approximates the B-WLS model. Thus, the B-GLS weights are almost equal to the B-WLS weights causing the RMSD to be very small. Also the LE from not using B-WLS is very small, which is due to the fact the B-GLS analysis only accounts for a minimal amount of cross-correlation among the annual peak floods. By comparing the RMSD and LE values for the California regional skew study to those from the Southeast U.S. regional skew study, thresholds of the metrics can be anticipated which will indicate when a B-GLS analysis might fail. Since the B-GLS methodology was successfully used to generate a regional skew model in the Southeast U.S., it seems that the thresholds of RMSD and LE should be at least as large as those values found in the Southeast U.S. By inspecting Figure 3.12a, it seems that the B-GLS weights produced by using cross-correlation Model 2 are still relatively noisy and negative, especially when compared to those weights depicted for the Southeast U.S. in Figure 3.13. Thus it seems that the thresholds for RMSD and LE should be less than the values calculated for Model 2. Thus, perhaps B-GLS could be considered likely to fail if RMSD > 0.7 and/or LE > 20%. While

these metrics are useful in helping to identify when B-GLS might fail, it is also important to inspect the weights.

3.4.2 Alternative Methodology

Do to the large cross-correlations among the annual peak floods and the results from Section 3.3.1, an alternative procedure is developed so that the regional skew analysis will provide stable and defensible results. To this end, a (Bayesian) Weighted Least Squares analysis is first used to develop estimators of the regression coefficients for each regional skew model. By using WLS, the cross-correlations are not employed in estimating the regression coefficients. After the regression model coefficients are determined with WLS, the precision of a model and the precision of the regression coefficients are estimated using a modified GLS analysis. However, due to the extensive use of low outlier screening and inclusion of historical information in the EMA analysis used in California, the simple formulas provided in Bulletin 17B and Griffis and Stedinger [2009] do not represent the variance of these sample skewness estimators. Thus, a Monte Carlo study was conducted to determine the actual sample variance of the skewness coefficient when a low outlier test is employed to identify samples for special treatment. Finally, a modified Bayesian GLS analysis using only data from pristine sites (i.e. sites without low outliers, zero flows, reconstructed records, or historical information) provided the estimate of the model error variance (the precision of the model) and the precision of the estimated parameters.

The specific computational steps employed in the California regional skew analysis are described below.

3.4.2.1 WLS Analysis to Generate Weights and Regression Parameters

A Weighted Least Squares (WLS) analysis is used to derive the regression model parameter estimates using the complete set of records. The resultant model yields an unbiased regional estimator of the skew at any site. The WLS analysis explicitly reflects variations in record length, but does not use estimates of cross correlation.

The WLS analysis is performed in two steps using unbiased at-site sample skewness estimators when possible. First, Bayesian-Weighted Least Squares (B-WLS) is used to estimate of the model error variance, denoted $\sigma^2_{\delta,B-WLS}$. Then, using $\sigma^2_{\delta,B-WLS}$, a Method-of-Moments WLS (MM-WLS) analysis is used to generate the weights, **W**, needed to estimate the regression parameters $\hat{\boldsymbol{\beta}}$. In order to compute the MM-WLS weights, a diagonal covariance matrix, $\Lambda_{WLS}\left(\sigma^2_{\delta,B-WLS}\right)$, is created. The diagonal elements of the covariance matrix are the sum of both the estimated model error variance, $\sigma^2_{\delta,B-WLS}$, and the variance of the unbiased at-site skew, $Var\left[\hat{\gamma}_i\right]$, which depends upon the record length N_i at each site. The unbiased at-site skew, $\hat{\gamma}_i$, used in the WLS analysis, as well as the variance of the at-site skew estimator, $Var\left[\hat{\gamma}_i\right]$, only depend on the annual peak flow record at each site *i*, and are calculated using the Expected Moments Algorithm (EMA) developed by Cohn *et al.* [1997]. The off-diagonal elements of $\Lambda_{WLS}\left(\sigma^2_{\delta,B-WLS}\right)$ are all zero, because cross-correlations between gage sites are not considered in a WLS anlaysis. Thus the (*n* x *n*) covariance matrix, $\Lambda_{WLS}\left(\sigma^2_{\delta,B-WLS}\right)$, is,

$$\boldsymbol{\Lambda}_{WLS}\left(\sigma^2_{\delta,B-WLS}\right) = \sigma^2_{\delta,B-WLS}\mathbf{I} + diag\left(Var[\hat{\gamma}]\right) \tag{3.14}$$

where *n* is the number of gage sites in the study, **I** is an (*n* x *n*) identity matrix, and $diag\left(Var[\hat{\gamma}]\right)$ is an *(n* x *n*) matrix containing the variance of the unbiased at-site sample

skewness estimators, $Var\left[\hat{\gamma}_i\right]$, on the diagonal and zeros on the off-diagonal. Using that covariance matrix, the MM-WLS weights are calculated as

$$\mathbf{W} = \left[\mathbf{X}^T \boldsymbol{\Lambda}_{WLS} \left(\sigma^2_{\delta,B-WLS}\right)^{-1} \mathbf{X}\right]^{-1} \mathbf{X}^T \boldsymbol{\Lambda}_{WLS} \left(\sigma^2_{\delta,B-WLS}\right)^{-1} \tag{3.15}$$

where, $\mathbf{W}$ is (k x n) matrix of weights and X is an (n x k) matrix of basin parameters, and k is the number of basin characteristics. These weights are used to compute the final estimates the regression parameters $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}} = \mathbf{W}\hat{\boldsymbol{\gamma}} \tag{3.16}$$

where $\hat{\boldsymbol{\beta}}$ is an (k x 1) vector of regression parameters and $\hat{\boldsymbol{\gamma}}$ is an (n x 1) vector of unbiased at-site sample skewness estimators. Both the B-WLS and MM-WLS computation include sites with historical information, zero flows, and low outliers.

3.4.2.2 Monte Carlo Analysis to Adjust for Bias in Pristine Data Set due to Lack of Low Outliers

After estimating the regression parameters, the true model error variance needs to be estimated. However, the extensive censoring of low outliers, the occurrence of zero flows, and the addition of regional historical flood information at some sites in the data set complicated the estimation of the model error variance. Thus, for the purpose of estimating the true model error variance, a simpler, "pristine" data set is developed. The pristine data set is a subset of the larger data set used in the WLS estimation of the regression parameters, and it does not include sites with historical information, zero flows, or low outliers as determined by EMA, or any reconstructed flow records.

Due to the exclusion of sites with low outliers in the pristine data set, the formulas provided in Bulletin 17B and Griffis and Stedinger [2009] misrepresent the variance of the

sample skewness estimators. Thus, a new Monte Carlo study is conducted to determine the actual variance of the skewness coefficient when a low outlier test is employed to identify low outliers. The Monte Carlo analysis of sample skews from a LP3 distribution retained only complete samples which did not contain low outliers. It is used to determine the bias associated with the sample skewness coefficient G (G is the traditional biased sample skewness coefficient) when samples with low outliers are dropped from the analysis. Two functions were computed: the mean of the sample skew, denoted $m(\gamma, N)$, as well as its variance, denoted $v(\gamma, N)$,

$$m(\gamma, N) = E_{\{x_i | no-outliers\}}[G \mid \gamma, N] \tag{3.17}$$

$$v(\gamma, N) = Var_{\{x_i | no-outliers\}}[G \mid \gamma, N] \tag{3.18}$$

These expectations are computed over only those LP3 samples, $\{x_i\}$, which do not contain low outliers, as determined by a 10% Grubbs-Beck test recommended by Bulletin 17B [IACWD, 1982].

Figure 3.14 shows the Monte Carlo results for $N = 50$ years of at-site annual peak flows. The x-axis is the true/population skew γ and the y-axis represents both the mean of the estimated skew $m(\gamma, N)$ and the standard deviation of the estimated skew (the square root of the variance $v(\gamma, N)$) across LP3 samples which had no low outliers. Figure 3.14 shows that when only samples without low outliers are considered, significant bias can be present in the mean of the estimated skews (the dashed line). Samples with low outliers are very likely to have negative skews, so starting with an unbiased estimator and omitting samples with low outliers is expected to yield a regional skewness estimator with a positive bias, if no correction is made.

When the true skew is highly positive, the bias is small and slightly negative because relatively few samples are omitted due to the presence of low outliers. When the true skew is

highly negative, the sample standard deviation is greatly reduced, as shown by the dotted line in Figure 3.14.

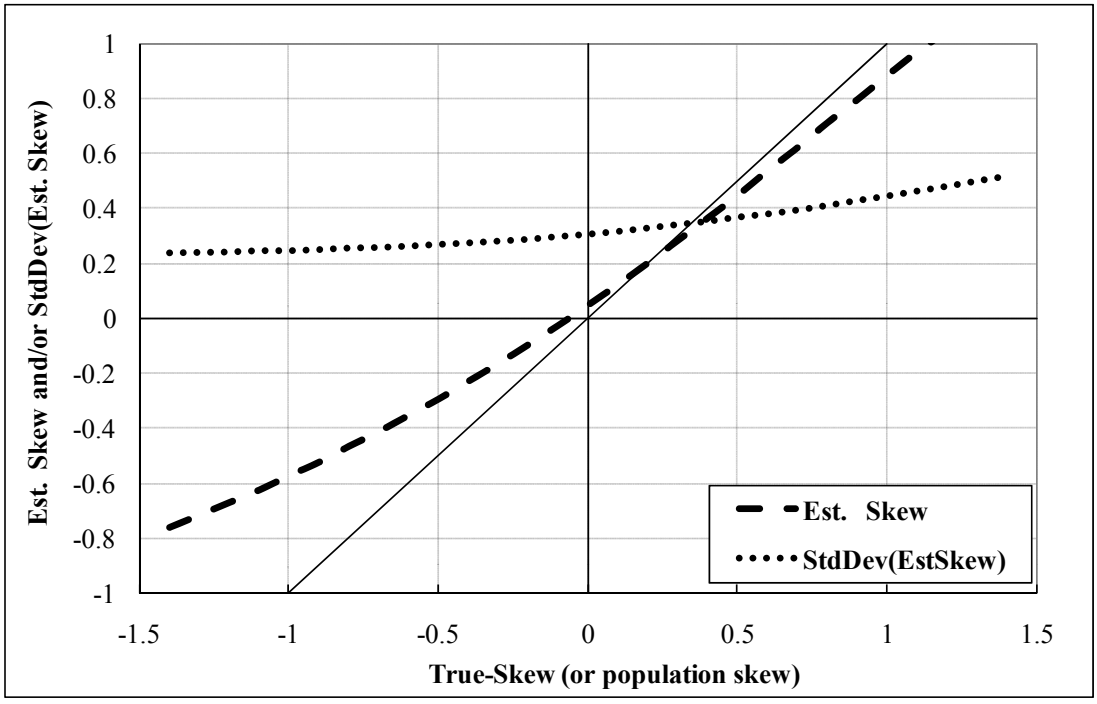


Figure 3.14: Graph of Monte Carlo results for N=50 years of at-site annual peak flows. The dashed line represents the mean $m(\gamma, N)$ of the estimated skew across samples without outliers and the dotted line represents the standard deviation (the square root of the variance $v(\gamma, N)$) of that estimated skew across Person type 3 samples which do not contain low outliers.

As shown in Table 3.7, the large negative skews had the highest percentage of samples with low outliers. For example, with a true at-site skew of -0.5, some 32% of the samples had low outliers and subsequently were dropped, as compared to a true at-site skew of 0.0 in which 10% of the samples were dropped. On the other hand, when the true at-site skew is highly positive, relatively few samples are rejected, resulting in very little impact on the sampling distribution of the estimated skew. In the Monte Carlo simulation, when the at-site record length is 50 years and the skew is 0.4, only 1% of the samples contained a low outlier. Table 3.7 reports that for populations with a skew of 0.5 or greater, no low outliers were observed in any of the

50,000 samples generated. It is easy to see why the standard deviation of the estimated skew increases as the value of the true skew increases.

Table 3.7: Monte Carlo results showing the percent of samples, for different at-site log-skew values, dropped from the simulation due to the presence of low outliers

Skew:	-1.5	-1.0	-0.8	-0.5	-0.3	0.0	0.3	0.5	0.8	1.0	1.5
% of Samples Dropped:	71%	56%	47%	32%	22%	10%	2%	0%	0%	0%	0%

3.4.2.3 Model Error Variance Estimation Using Bayesian GLS

The Monte Carlo experiments provide the expected value and variance of the sample skewness estimator from samples without outliers. These functions can then be used with the pristine data set in the computations of the model error variance employing a B-GLS framework. By using the relationship for regional skew generated by the WLS analysis, the WLS mean regional skew estimate $m_R(i)$, can be calculated for each site i in the pristine data set as:

$$\mathbf{m}_R = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{3.19}$$

where $\mathbf{m}_R$ is an (n_p x 1) vector of WLS regional skew estimates for each site in the pristine data set, $\mathbf{X}$ is an (n_p x k) matrix of basin parameters, $\hat{\boldsymbol{\beta}}$ is an (k x 1) vector of WLS regression parameters, n_p is the number of gage sites in the pristine data set, and k is the number of basin characteristics.

The last step is to estimate the model error variance using the pristine data set. If the model error variance were zero, $\sigma^2_{\delta,B-GLS} = 0$, then all of the observed variability would be sampling error, and we would have

$$E\left[\left\{G_i - m(\gamma_i, N_i)\right\}^2\right] = v(\gamma_i, N_i) \tag{3.20}$$

However, we anticipate that the model will not be perfect, and thus estimation of the model error variance will be more challenging. The derivative

$$d\left\{m\left(\gamma,N\right)\right\}/d\gamma \neq 1 \tag{3.21}$$

will be used in a correction to our GLS analysis. For at a given site with record length N, let

$$r = d\left\{m\left(\gamma,N\right)\right\}/d\gamma \tag{3.22}$$

and let $\sigma^2_{\delta,B-GLS}$ be the model error variance. Then to first order:

$$E\left[\left\{G_i - m\left(\gamma_i,N_i\right)\right\}^2\right] = r^2\sigma^2_{\delta,B-GLS} + v\left(\gamma_i,N_i\right) \tag{3.23}$$

Thus, the GLS covariance matrix for the pristine data set is:

$$\mathbf{\Lambda_p}\left(\sigma^2_{\delta,B-GLS}\right) = r^2\sigma^2_{\delta,B-GLS}\mathbf{I_p} + \mathbf{\Sigma}\left(G\right) \tag{3.24}$$

where $\mathbf{\Lambda_p}\left(\sigma^2_{\delta,B-GLS}\right)$ is an (n_p x n_p) GLS covariance matrix, $\mathbf{I_p}$ is an (n_p x n_p) identity matrix, $\mathbf{\Sigma}\left(G\right)$ is an (n_p x n_p) matrix containing the sampling variances of the biased skewness estimators $Var\left[G_i\right] = v\left(\gamma_i,N_i\right)^2$ and the covariances of the skewness estimators G_i in the pristine data set. The value of $\mathbf{\Sigma}\left(G\right)$ is determined by the length of record at each station and the cross-correlation of the concurrent flows.

The covariance matrix for the skewness coefficients for the pristine data set, $\Lambda_p\left(\sigma^2_{\delta,B-GLS}\right)$, and the conditional means of the sample skews, $m\left(\gamma,N\right)$, are used in a Bayesian framework to compute the posterior distribution of the model error variance, and in particular the posterior mean of the true model error variance.

3.4.2.4 Estimation of Regression Parameter Precision

The B-GLS model error variance can then be used to compute the precision of the regression parameters $\hat{\boldsymbol{\beta}}$ that were calculated with the WLS weights $\mathbf{W}$, as $\hat{\boldsymbol{\beta}} = \mathbf{W}\hat{\boldsymbol{\gamma}}$. The variance of $\hat{\boldsymbol{\beta}}$ is simply

$$Var\left[\hat{\boldsymbol{\beta}}\right] = \mathbf{W}\boldsymbol{\Lambda}\left(\sigma^2_{\delta,B-GLS}\right)\mathbf{W}^T \tag{3.25}$$

where $\boldsymbol{\Lambda}\left(\sigma^2_{\delta,B-GLS}\right)$ is an (n x n) covariance matrix which uses all of the sites, not just those in the pristine data set, and $\sigma^2_{\delta,B-GLS}$ corresponds to the model error variance calculated from the B-GLS analysis described above. It is important to note that $\boldsymbol{\Lambda}\left(\sigma^2_{\delta,B-GLS}\right)$ contains the B-GLS model error variance and thus is not the same as the covariance matrix $\boldsymbol{\Lambda}_{WLS}\left(\sigma^2_{\delta,B-WLS}\right)$ used in the MM-WLS analysis of all of the sites which contains the B-WLS model error variance. Also, $\boldsymbol{\Lambda}\left(\sigma^2_{\delta,B-GLS}\right)$ is an (n x n) covariance matrix which uses all of the sites, and thus is not the same as the (n_p x n_p) covariance matrix $\boldsymbol{\Lambda}_p\left(\sigma^2_{\delta,B-GLS}\right)$ used in the B-GLS analysis of just the pristine sites.

3.5 California Regional Skewness Analysis

The Southeastern United States regional skew analysis illustrates how a Bayesian GLS analysis would generally proceed [Veilleux, 2009; Weaver *et al*., 2009; Feaster *et al*., 2009; and Gotvald *et al*., 2009]. However, when a Bayesian GLS analysis was attempted with the California data set, consistent results were not obtained due to the large cross-correlations. Thus, an alternative procedure which uses both Weighted Least Squares (WLS) and B-GLS was developed so that the regional skew analysis would provide more stable and defensible results. The needs for the alternative procedure, as well as the specific computational steps are described

in Section 3.3. The results of the California regional skew regression using those procedures are provided below.

All of the available basin characteristics were initially considered as explanatory variables in the regression regional skew analysis. The one key basin characteristic that was statistically significant in explaining the site-to-site variability in skew was the Mean Basin Elevation (ELEV). Table 3.8 presents the final results for three models: a constant skew denoted "Constant," a model that uses a linear relationship between skew and basin elevation denoted "Elev,", and a model that uses a nonlinear relationship between skew and mean basin elevation denoted : "NL-Elev."

Table 3.8: Regional skew models produced by extended GLS analysis of California annual maximum flows. Standard deviations are presented in ( ), σ_δ^2 is the model error variance, ASEV is the average sampling error variance, AVPnew is the average variance of prediction for a new site.

Model	Regression Parameters B0	B1	B2	σ_δ^2	ASEV	AVP_{new}	R_δ^2
Constant: $\hat{\gamma} = \beta_0$	-0.23 (0.03)	-	-	0.20 (0.06)	0.03	0.23	0%
Elev: $\hat{\gamma} = \beta_0 + \beta_1 (Elev)$	-0.76 (0.05)	1.4E-04 (1.2E-09)	-	0.12 (0.04)	0.03	0.15	41%
NL Elev: $\hat{\gamma} = \beta_0 + \beta_2 \left[1 - \exp\left(-\left(Elev/6500\right)^2\right)\right]$	-0.62 (0.04)	-	1.30 (0.10)	0.10 (0.04)	0.03	0.14	48%

As shown in Table 3.8, the linear "Elev" model has a pseudo R_δ^2 of 41%, while the nonlinear "NL-Elev" model has a larger pseudo R_δ^2 of 48%, and a smaller AVP_{new}. The pseudo R_δ^2 values describe the fraction of the variability in the true skews explained by each model [Gruber and others, 2007]. A constant model does not explain any variability, so the pseudo R_δ^2

is equal to 0%. As shown in Table 3.8, the posterior mean of the model error variance, σ_δ^2, for the "NL-Elev" is 0.10, which is smaller than that for both the linear "Elev" model (σ_δ^2=0.12) and is half the value for the constant model (σ_δ^2=0.20). The Average Sampling Error Variance (ASEV) presented in Table 3.8 accounts for the average error in the estimator of the skew at the sites in the data set.

The Average Variance of Prediction at a new site (AVP_{new}) corresponds to the MSE employed in Bulletin 17B to describe the precision of the generalized skew. In Table 3.8, the "NL-Elev" model has the lowest AVP_{new} equal to 0.14. However, this AVP_{new} is an average value computed by averaging the Variance of Prediction at a new site (VP_{new}) over all of the 158 sites in the California study. Just as generalized skew varies from site-to-site depending upon ELEV, so too does the value of VP_{new}. Section 3.5.1 provides a detailed discussion of VP_{new} for the California study.

Figure 3.15 is a graph of the unbiased at-site skewness versus mean basin elevation in feet. The 158 gauge sites used to develop the regional skew models are plotted as circles, while the constant model is plotted as solid black line, the "Elev" model is plotted as •'s, and the "NL-Elev" model is plotted as +'s.

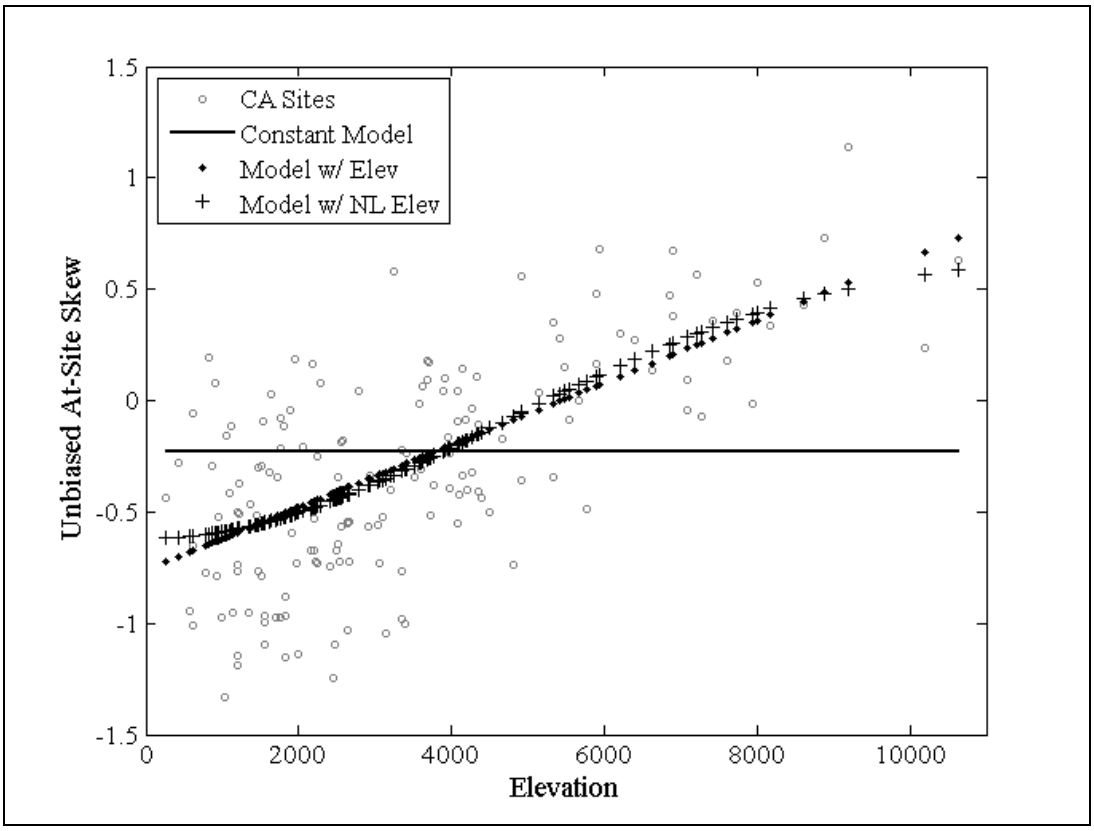


Figure 3.15: Relationship between unbiased at-site skew and mean basin elevation for 158 sites in California. The solid black line represents the constant model from Table 3.8, the •'s represent the "Elev" model from Table 3.8, and the +'s represent the "NL-Elev" model from Table 3.8. The open circles are the 158 gauge sites used to construct the models in Table 3.8.

As shown in Figure 3.15, the nonlinear elevation model provides a reasonable fit for the California regional skew data. While the more complicated nonlinear model is not very different from the simpler linear model, the nonlinear model provides smaller values of positive skew at high elevations and less negative values of skew for low elevations. For example, when mean basin elevation is zero at sea level, the nonlinear model provides a regional skew of –0.62, while the linear elevation model provides a regional skew of –0.76. Conversely, at a mean basin elevation of 11,000 ft, the nonlinear model provides a regional skew of 0.61, while the linear model provides a regional skew 0.79. These differences, though subtle, are significant, and the nonlinear model indicates that regional skew flattens out in the tails instead of continually

increasing in absolute value. This flattening of skew at both low and high elevations is consistent with the relation between timing of annual peak discharge and elevation, which is reflective of the degree of rain-snow interaction effects on peak discharge. Annual peak-discharges from basins with mean elevations less than about 4,000 feet have little rain-snow interaction, and therefore might be expected to have constant or near-constant regional skews. Likewise, basins at very high elevation tend to have annual peak discharges series dominated by snow events, and thus show less of an elevation effect on regional skew. See Section 3.2 for detailed discussion of California hydrology and the effect of the rain-snow interaction on regional skewness.

As indicated in Figure 3.15, only a few sites with mean basin elevation greater than about 8,000 *ft* were used in the regional skew analysis. Because of the scarcity of data for such high-elevation sites, coupled with the greater effects of rain-snow interaction at these high elevations, the calculated regional skew values for high-elevation sites may be less reliable and the ability of the log Pearson Type-3 (LP3) distribution to describe these series is a concern. Peak-discharge data for sites with mean basin elevations greater than about 8,000 *ft* should be examined to determine if mixed-population methods for determining flood frequency described in Bulletin 17B might be more appropriate than the standard application of the LP3 method.

3.6 California Bayesian GLS Regression Diagnostics

The goal of the regression diagnostics is to allow for a comprehensive examination of a regression analysis developed within the extended B-GLS framework described in Section 3.3.2. Sections 3.5.1 and 3.5.2 discuss the Variance of Prediction and pseudo Analysis of Variance Table, respectively, for the California regional skewness models presented in Section 3.4.

3.6.1 Variance of Prediction

Variance of Prediction is a common metric used to choose which of several models provides the most accurate estimator of the *y*-variable, because it combines both the model error variance and the sampling error in the model parameters. The Variance of Prediction at a new site *i* is

$$VP_{new}(i) = \underset{\sigma^2_{\delta,B-GLS}}{E}\left[\sigma^2_{\delta,B-GLS} + \mathbf{x}_i\mathbf{X}\mathbf{\Lambda}\left(\sigma^2_{\delta,B-GLS}\right)\mathbf{X}^T\mathbf{x}_i^T\right] \tag{3.26}$$

where $\mathbf{X}$ is an (n x k) matrix of basin parameters, n is the number of gage sites in the study, k is the number of basin characteristics, and $\mathbf{\Lambda}\left(\sigma^2_{\delta,B-GLS}\right)$ is the (n x n) covariance matrix.

As stated in Section 3.4, the Average Variance of Prediction at a new site (AVP_{new}) corresponds to the MSE employed in Bulletin 17B to describe the precision of the generalized skew. The AVP_{new} is an average value computed by averaging the Variance of Prediction at a new site (VP_{new}) over all of the 158 sites in the California study. Just as generalized skew varies from site-to-site depending upon ELEV, so too does the value of VP_{new}. Table 3.9 presents values of regional skew, VP_{new}, and Effective Record Length (ERL) for the "NL-Elev" model for various values of the mean basin elevation ELEV between 0 and 11,000 feet. Figure 3.16 below illustrates how VP_{new} and ERL vary as mean basin elevation increases. The ERL is a function of VP_{new} and the average regional skew. The ERL is proportional to the average regional skew and inversely proportional the VP_{new}. (See Griffis and Stedinger, 2009; eqns 3-9).

Table 3.9: Variance of Prediction (VP) and Equivalent Record Length (ERL) for "NL-Elev" model for various values of Mean Basin Elevation (ELEV).

Elevation (ft)	Avg. Regional	VP_{new}	ERL
0	-0.62	0.143	65
1000	-0.59	0.141	65
2000	-0.50	0.138	62
3000	-0.37	0.134	58
4000	-0.21	0.132	55
5000	-0.04	0.133	53
6000	0.13	0.137	52
7000	0.28	0.144	52
8000	0.40	0.151	53
9000	0.49	0.158	54
10000	0.56	0.163	55
11000	0.61	0.168	55

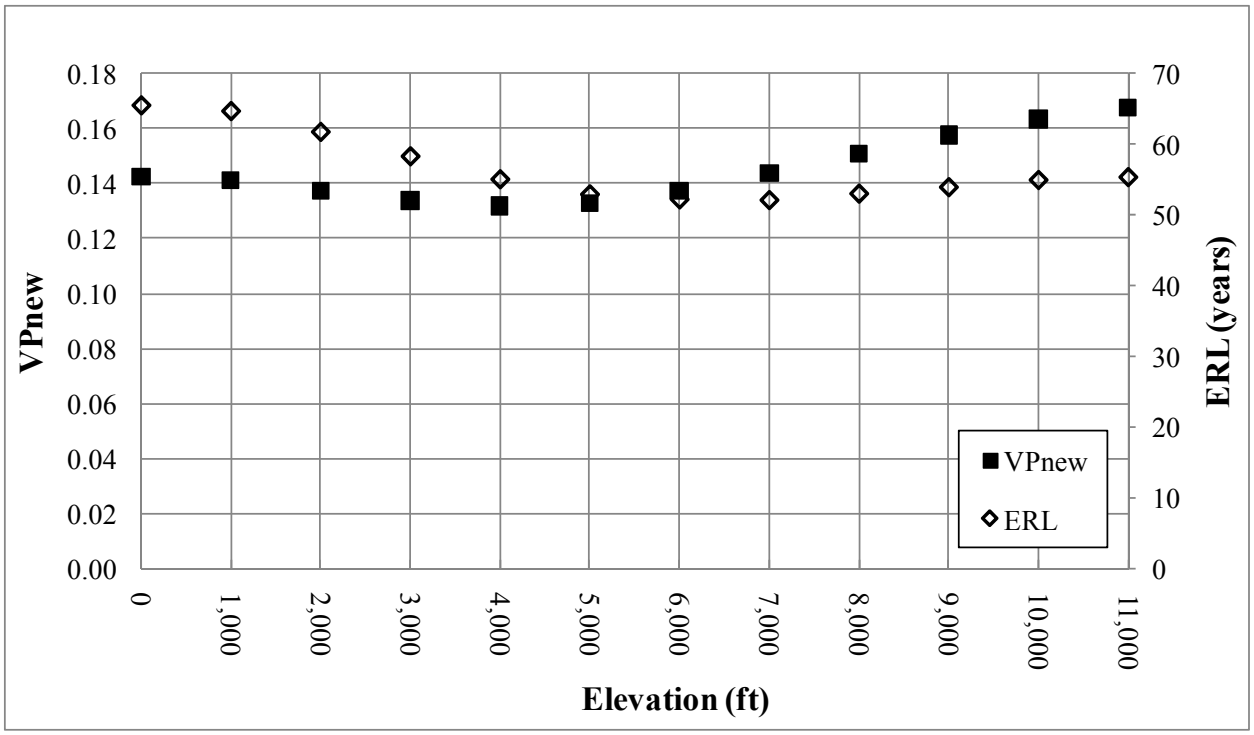


Figure 3.16: Variance of Prediction at a new site (VPnew) and Effective Record Length (ERL) as a function of mean basin elevation in feet for the "NL-Elev" model from Table 3.9. The solid squares represent the VPnew values corresponding to the left-hand *y*-axis, while the open diamonds represent the ERL corresponding to the right-hand *y*-axis.

The "NL-Elev" regional skew model for California has an effective record length between 52 years and 65 years, depending on mean basin elevation. A VP ranging from about 0.13 to 0.17 is a marked improvement over the Bulletin 17B skew map, whose MSE is 0.302 [IACWD, 1982] with a corresponding effective record length of only 17 years.

3.6.2 Pseudo ANOVA

To determine if a model is a good representation of the data and which regression parameters, if any, should be included in a regression model, diagnostic statistics have been developed to evaluate how well a model fits a regional hydrologic data set [Griffis and Stedinger, 2007; Gruber and Stedinger, 2008]. Here the goal of model selection is to resolve which set of possible explanatory variables best fit the California flood data affording the most accurate skew prediction, while also allowing for the simplest model possible. This section presents the diagnostic statistics for a Bayesian WLS or GLS analysis, and discusses the specific values obtained for the California regional skew study.

Table 3.10 presents a Pseudo Analysis of Variance (Pseudo ANOVA) table for the California regional skew analysis. The table contains regression diagnostics/goodness of fit statistics which are explained below.

Table 3.10. Pseudo ANOVA table for the California regional skew study for both the Constant Model and the NL-Elev Model

Source	Degrees-of-Freedom			Equations	Sum of Squares	
		Constant	NL-Elev		Constant	NL-Elev
Model	***k***	0	1	$n\left[\sigma_\delta^2(0)-\sigma_\delta^2(k)\right]$	0.0	15
Model Error	***n-k-1***	157	156	$n\sigma_\delta^2(k)$	32	17
Sampling Error	***n***	158	158	$\sum_{i=1}^{n} Var(\hat{\gamma}_i)$	34	34
Total	***2n-1***	315	315	$n\sigma_\delta^2(0)+\sum_{i=1}^{n} Var(\hat{\gamma}_i)$	66	66
EVR					1.1	2.1
MBV*					13	16
$\mathbf{R_\delta^2}$					0%	48%

In particular, Table 3.10 describes how much of the variation in the observations can be attributed to the regional model, and how much of the residual variation can be attributed to model error and sampling error, respectively. Difficulties arise in determining these quantities. The model errors cannot be resolved because the values of the sampling errors η_i for each site *i,* are not known. However, the total sampling error sum of squares can be described by its mean value, $\sum_{i=1}^{n} Var[\hat{\gamma}_i]$. Because there are *n* equations, the total variation due to the model error δ for a model with *k* parameters has a mean equal to $n\sigma_\delta^2(k)$. Thus, the residual variation attributed to the sampling error is $\sum_{i=1}^{n} Var[\hat{\gamma}_i]$, and the residual variation attributed to the model error is $n\sigma_\delta^2(k)$.

For a model with no parameters other than the mean (*i.e* the constant skew model), the estimated model error varaince $\sigma_\delta^2(0)$ describes all of the anticipated variation in $\gamma_i = \mu + \delta_i$, where μ is the mean of the estimated at-site sample skews. Thus, the TOTAL expected sum of squares variation due to model error δ_i and due to sampling error $\eta_i = \hat{\gamma}_i - \gamma_i$ in expectation should equal $n\sigma_\delta^2(0) + \sum_{i=1}^{n} Var(\hat{\gamma}_i)$. Therefore, the expected sum of squares attributed to a regional skew model with k parameters equals $n[\sigma_\delta^2(0) - \sigma_\delta^2(k)]$, because the sum of the model error variance $n\sigma_\delta^2(k)$ and the variance explained by the model must sum to $n\sigma_\delta^2(0)$. Table 3.10 considers models with k = 0 and 1.

This division of the variation in the observations is referred to as a Pseudo ANOVA because the contributions of the three sources of error are estimated or constructed, rather than being determined from the computed residual errors and the observed model predictions, while also ignoring the impact of correlation among the sampling errors.

Table 3.10 compares the Pseudo ANOVA results for the Constant model and NL-ELEV model. Both models have the same sampling error because they both use the same set of at-site skew data. Both the Constant model and the NL-ELEV model have sampling error variances larger than their model error variances. However, it is important to note that the model error variance attributed to the NL-ELEV model $\sigma_\delta^2(1)$ is almost half of the model error variance for the Constant Model $\sigma_\delta^2(0)$. This difference in model error is accounted for by the variation in the sample that the NL-model model appears to explain. Because the Constant model does not have any explanatory variables, the variation attributed to that model is 0. On the other hand, the NL-ELEV model has one explanatory variable, which causes the variation attributed to the

resultant model to increase to 15. This reduces the model error variance from 32 with the Constant model to 17 with the NL-ELEV model; thus the addition of the nonlinear elevation explanatory variable in the NL-ELEV model greatly improves the ability of the model to describe the observed skew coefficients. This impact is described by the pseudo R_{δ}^{2}, which in this case has a value of 48% because the NL_ELEV model explains 48% of the estimated variation $\sigma_{\delta}^{2}(0)$ in the true skew from site-to-site.

The Pseudo Analysis of Variance also provides the information needed to evaluate if a sophisticated WLS or GLS analysis is needed to correctly interpret the data. In particular, the Error Variance Ratio (EVR) is a modeling diagnostic used to evaluate if a simple OLS regression is sufficient or a more sophisticated WLS or GLS analysis is appropriate. EVR is the ratio of the average sampling error variance to the model error variance. Generally, an EVR greater than 20%, indicates that the sampling variance is not negligible when compared to the model error variance, suggesting the need for a WLS or GLS regression analysis. The EVR is calculated as

$$EVR = \frac{\mathrm{SS}(\text{sampling error})}{\mathrm{SS}(\text{model error})} = \frac{\sum_{i=1}^{n} Var(\hat{\gamma}_i)}{n\sigma_{\delta}^{2}(k)} \tag{3.27}$$

In this case, EVR had a value of 1.1 for the constant model, and 2.1 for the nonlinear-elevation model. The sampling variability in the sample skewness estimators was as large or larger than the error in the regional model. Thus, most likely, given the variation of record lengths from site-to-site, it is important to use a WLS or GLS analysis to evaluate the final precision of the model, rather than a simpler analysis that neglected the sampling error in the at-site skewness estimators.

The Misrepresentation of the Beta Variance (MBV) statistic is used to determine whether a WLS regression is sufficient, or if a GLS regression is needed to determine the precision of the estimated regression parameters [Griffis and Stedinger, 2007]. The MBV describes the error produced by a WLS regression analysis in its evaluation of the precision of b_0^{WLS}, which is the estimator of the constant β_0^{WLS}, because the covariance among the estimated at-site skews $\hat{\gamma}_i$ generally has its greatest impact on the precision of the constant term [Stedinger and Tasker, 1985]. If the MBV is substantially greater than 1, then a GLS error analysis should be employed. The MBV* (the updated version of the MBVrecommended in Griffis and Stedinger [2007] as discussed in Section 2.8) is calculated as,

$$MBV^* = \frac{Var\left[b_0^{WLS} \mid GLS analysis\right]}{Var\left[b_0^{WLS} \mid WLS analysis\right]} = \frac{w^T \Lambda w}{\sum_{i=1}^{n} w_i} \qquad \text{where } w_i = \frac{1}{\Lambda_{ii}} \tag{3.28}$$

For the California regional skew study, the MBV* is equal to 16 for the NL-Elev model, and 13 for the Constant model. This is a very large value indicating the cross-correlation among the skewness estimators has had a major impact on the precision with which the regional average skew coefficient can be estimated; if a WLS precision analysis were used for the estimated constant parameter in the NL_ELEV model, the variance would be underestimated by a factor of 16. Thus a WLS analysis would have seriously misrepresented the variance of the constant in the Constant Model and in the NL-ELEV model of regional skew; this would have resulted in underestimation estimation of the variance of prediction given that the sampling error in the constant term in both models was sufficiently large enough to make an appreciable contribution to the average variance of prediction.

3.7 Conclusions

Based upon the regional skew analysis of the selected California stations, the recommended regional model is

$$\hat{\gamma} = \beta_0 + \beta_2 \left[1 - \exp\left(-\left(Elev/6500 \right)^2 \right) \right] \tag{3.29}$$

with MSE = 0.14. The constant model had a MSE of 0.23. Either is a definite improvement over the Bulletin 17B skew map which reports a MSE of 0.302. Much of the difference occurs because the new analysis correctly reflects the difference between the sampling error in at-site skew coefficient estimators, and the precision of the regional model [Hardison, 1971; Stedinger and Tasker, 1985; and Tasker and Stedinger, 1986].

The extended Bayesian-GLS methodology developed in this chapter provides stable and defensible results for the California regional skew analysis. The Bayesian-WLS analysis first developed estimators of the regression coefficients for each regional skew model. By using WLS, the cross-correlations are not employed in estimating the regression coefficients. After the regression model coefficients were determined with WLS, the precision of a model and the precision of the regression coefficients are estimated using a modified GLS analysis. A Monte Carlo analysis determined the actual sample variance of the skewness coefficient when a low outlier test is employed to identify samples for special treatment. Finally, a modified Bayesian GLS analysis using only data from pristine sites (i.e. sites without low outliers, zero flows, reconstructed records, or historical information) provided the estimate of the model error variance (the precision of the model) and the precision of the estimated parameters. This extended Bayesian-GLS methodology provided a stable regional skew model for California, while avoiding the instability issues encountered by the original B-GLS methodology due to the large cross-correlations between annual peak floods.

It is interesting that the Bulletin 17B skew map had iso-skew lines that assigned a negative -0.30 skew to basins along the coast, and positive +0.2 skews to basins along the ridge of the Sierra. An iso-line for zero skew ran along the center of the Central Valley. The new function has a similar trend, but ties the differences in regional skew to the specific mean elevation of each basin. Elevation is a good physical indicator of the character of each basin and the relative importance of rain and snow events, which explains the observed differences in station skew. While having the same overall trend, in many cases the new function will provide very different estimates of regional skew.

The hydrology of California confirms the results presented in the statistical model; the annual maximum flood flows in California are realted to the mean basin elevations through a nonlinear function. It is shown that gauge sites with mean basin elevation below 4,000 *ft* have their maximum annual floods driven by a different hydrologic mechanism than those basins with elevations above 4,000 *ft*. Those sites with mean basin elevations below 4,000 *ft* have their maximum annual floods driven by rainfall events. However as the mean elevation of basins increases above 4,000 *ft,* the interaction of rainfall and snowmelt events increasingly effects the maximum annual floods. As shown in the regional skew model developed in Section 3.5, this hydrology, described through a nonlinear model of mean basin elevation, helps significantly to explain regional skewness in California.

APPENDIX A

CALIFORNIA STREAM FLOW GAUGE SITES

This appendix contains the 192 gauge sites used in the California regional flood skewness estimate. Table A contains the USGS site number, site index number, years of record, the EMA estimated at-site log skewness the variance of that skewness, centroid location, drainage basin area, and mean basin elevation for each of the 192 gauge stations in the study.

Table A: The 192 peak stream flow gauge sites and their basin characteristics used in the California regional skew study

USGS Hydrologic Unit Code	Site Index #	Years of Record	EMA At-Site Log Skew	EMA At-Site Var(Log Skew)	Basin Centroid Location Lat	Basin Centroid Location Lon	Drainage Area (mi^2)	Mean Basin Elevation (ft)
09423350[E]	1	42	0.097	0.147	35.25	-115.30	0.84	174
10251300[E]	2	31	-0.033	0.183	36.40	-116.29	3090	2904
10257600[E]	3	37	0.061	0.165	34.09	-116.69	35.6	5710
10258000[E]	4	57	-0.021	0.107	33.79	-116.63	16.9	6834
10259000[E]	5	58	0.368	0.126	33.76	-116.60	8.65	4480
10260500[E]	6	87	0.039	0.474	34.29	-117.09	134	5861
10261000[E]	7	49	-0.111	0.104	34.28	-117.29	70.3	4258
10263500[E]	8	83	-0.009	0.074	34.38	-117.81	22.9	6349
10263900[E]	9	34	-0.633	0.266	34.34	-117.93	0.48	7591
10264000[E]	10	50	-0.464	0.124	34.39	-117.97	49.0	5508
10264600[E]	11	29	0.074	0.210	35.03	-118.43	15.8	5767
10265200[E]	12	53	-0.408	0.141	37.56	-118.87	18.2	10249
10265700[E]	13	52	-0.256	0.130	37.45	-118.73	35.8	10685
10267000[E]	14	58	-0.992	0.224	37.36	-118.71	36.4	10413
10268700[E]	15	49	0.286	0.140	37.42	-118.22	20.0	8788
10281800[E]	16	56	-0.567	0.153	36.77	-118.34	18.1	9832
10286000[E]	17	68	-0.566	0.150	36.46	-118.16	40.1	10035
10291500[P]	18	37	0.979	0.360	38.20	-119.42	43.8	9193
10295500[R]	19	53	0.539	0.141	38.29	-119.45	63.1	206
10296500[P]	20	81	0.398	0.090	38.33	-119.49	244	8610
10336610[P]	21	31	0.148	0.143	38.81	-120.02	53.8	7609
10336676[P]	22	34	-0.060	0.129	39.14	-120.21	9.53	7281
10336780[P]	23	45	-0.015	0.106	38.87	-119.94	36.7	7932
10343500[P]	24	53	-0.041	0.093	39.43	-120.27	10.6	7095
10358500[P]	25	44	-0.429	0.158	40.63	-120.84	90.0	5772
11015000	26	72	-0.098	0.099	32.90	-116.57	45.4	4361
11028500	27	67	-0.472	0.152	33.05	-116.94	57.6	1441
11033000	28	30	-0.415	0.198	33.32	-116.82	25.5	4503
11042400	29	49	0.081	0.128	33.42	-116.79	131	3680
11046100[P]	30	39	-0.049	0.143	33.36	-117.41	26.6	608

Note: A = U.S. Army Corp of Engineers Gage Site, E = Eastern Sierra Lahontan Region Gage Site (removed), L = Gage site with lack of LP3 fit (removed), P = Pristine Gage Site, R = Redundant Gage Site (removed)

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	Years of Record	EMA At-Site Log Skew	EMA At-Site Var(Log Skew)	Basin Centroid Location Lat	Basin Centroid Location Lon	Drainage Area (mi^2)	Mean Basin Elevation (ft)
11055500[P]	31	88	-0.221	0.074	34.17	-117.13	16.9	3982
11055800[P]	32	87	0.092	0.071	34.20	-117.18	19.6	3919
11058500[P]	33	82	0.061	0.071	34.21	-117.24	8.80	3629
11058600[P]	34	68	0.246	0.096	34.22	-117.28	4.65	3688
11063000[P]	35	58	0.037	0.103	34.33	-117.46	58.9	3897
11073470[P]	36	46	0.250	0.142	34.20	-117.62	9.68	5406
11075800[P]	37	45	-0.500	0.179	33.71	-117.58	13.0	2917
11084000[P]	38	45	-0.301	0.154	34.20	-117.91	6.64	2514
11096500	39	45	-0.369	0.250	34.32	-118.33	21.1	2459
11098000[P]	40	92	-0.292	0.073	34.25	-118.15	16.0	3598
11100000[P]	41	54	-0.013	0.115	34.21	-118.02	9.71	3571
11100500[P]	42	46	-0.306	0.151	34.20	-118.05	1.84	3513
11104000	43	49	-0.417	0.183	34.10	-118.60	18.0	1361
11105850[P]	44	42	-0.872	0.258	34.29	-118.71	70.6	1558
11107745[P]	45	31	-0.199	0.173	34.46	-118.21	157	3423
11110500	46	48	-0.165	0.133	34.47	-118.84	23.6	2563
11113000	47	68	-0.508	0.120	34.56	-119.11	251	4091
11113500	48	72	-0.207	0.108	34.46	-119.08	38.4	3356
11117500	49	34	-0.037	0.145	34.45	-119.21	51.2	1884
11117600	50	29	0.135	0.216	34.44	-119.40	13.2	2175
11119500	51	65	-0.545	0.131	34.44	-119.48	13.1	1911
11120500	52	65	-0.890	0.179	34.49	-119.82	5.51	1692
11124500	53	64	-0.700	0.159	34.66	-119.79	74.0	3355
11126500	54	33	-0.275	0.237	34.67	-119.99	55.8	1610
11132500[P]	55	66	-0.725	0.151	34.55	-120.36	47.1	920
11134800[P]	56	35	-0.250	0.179	34.60	-120.47	11.6	868
11136100[P]	57	46	0.173	0.134	34.77	-120.33	135	818
11136800	58	49	-0.893	0.179	34.89	-119.63	886	3388
11138000[P]	59	32	-0.474	0.247	35.16	-120.33	117	1427
11138500[R]	60	61	-0.484	0.135	34.81	-119.88	281	3263

Note: A = U.S. Army Corp of Engineers Gage Site, E = Eastern Sierra Lahontan Region Gage Site (removed), L = Gage site with lack of LP3 fit (removed), P = Pristine Gage Site, R = Redundant Gage Site (removed)

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	Years of Record	EMA At-Site Log Skew	EMA At-Site Var(Log Skew)	Basin Centroid Location Lat	Basin Centroid Location Lon	Drainage Area (mi^2)	Mean Basin Elevation (ft)
11139500[P]	61	43	-0.188	0.132	34.94	-120.22	28.7	1762
11140000	62	63	-0.657	0.341	34.84	-119.99	471	2655
11141280	63	39	-0.098	0.295	35.28	-120.53	20.9	1811
11143000	64	56	-0.539	0.153	36.24	-121.68	46.5	2552
11143200	65	49	0.069	0.369	36.40	-121.64	193	2280
11143500[P]	66	41	-0.585	0.212	35.29	-120.30	70.3	2211
11147070	67	33	-0.646	0.242	35.53	-120.83	18.2	1463
11147500	68	63	-0.716	0.161	35.44	-120.58	390	1518
11148500	69	52	-0.657	0.408	35.48	-120.16	922	1981
11148900	70	35	-0.436	0.249	35.81	-121.08	162	1224
11151300[P]	71	48	-0.182	0.136	36.26	-120.85	233	2061
11152000[P]	72	101	-0.633	0.091	36.23	-121.47	244	2494
11152540[P]	73	40	-0.326	0.177	36.54	-121.69	31.9	1212
11152600	74	36	-1.019	0.364	36.77	-121.55	36.7	1193
11156500[R]	75	68	-0.079	0.091	36.45	-120.93	249	2727
11157500	76	54	-0.607	0.149	36.72	-121.12	208	2156
11158699[A]	77	57	-0.478	0.152	36.61	-121.10	606	2198
11159000	78	67	-0.885	0.171	36.79	-121.28	1186	1556
11159200[P]	79	51	-0.873	0.226	37.02	-121.80	27.8	1001
11160000	80	56	-0.455	0.402	37.07	-121.92	40.2	1196
11160500[P]	81	70	-0.679	0.137	37.14	-122.09	106	1196
11162500	82	55	-0.857	0.366	37.26	-122.22	45.9	1137
11162570	83	32	-0.284	0.297	37.34	-122.29	50.9	1048
11162630	84	40	-0.672	0.235	37.49	-122.40	27.1	783
11164500	85	67	-0.483	0.153	37.39	-122.24	37.4	953
11169500[P]	86	73	-0.075	0.085	37.24	-122.07	9.22	1754
11176000	87	54	-0.384	0.116	37.54	-121.57	38.2	2492
11176400	88	43	-0.963	0.291	37.43	-121.55	130	2479
11182100	89	39	-0.872	0.315	37.96	-122.20	10.0	606
11182500[P]	90	54	-0.588	0.162	37.77	-122.00	5.89	620

Note: A = U.S. Army Corp of Engineers Gage Site, E = Eastern Sierra Lahontan Region Gage Site (removed), L = Gage site with lack of LP3 fit (removed), P = Pristine Gage Site, R = Redundant Gage Site (removed)

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	Years of Record	EMA At-Site Log Skew	EMA At-Site Var(Log Skew)	Basin Centroid Location Lat	Basin Centroid Location Lon	Drainage Area (mi^2)	Mean Basin Elevation (ft)
11189500[P]	91	73	0.364	0.096	36.06	-118.19	530	7726
11200800[P]	92	39	-0.339	0.181	35.91	-118.69	83.30	3986
11203500[P]	93	59	-0.081	0.106	36.19	-118.74	253	4186
11204500[P]	94	59	0.038	0.104	36.03	-118.71	109	4076
11209500[R]	95	52	-0.010	0.108	36.61	-118.88	129	4950
11210500[P]	96	58	0.617	0.149	36.50	-118.77	519	5933
11211300[P]	97	37	-0.466	0.161	36.57	-119.00	75.6	2667
11212000	98	32	-0.250	0.217	36.67	-119.20	31.6	1516
11213500[R]	99	53	0.643	0.170	36.86	-118.68	952	8593
11214000[P]	100	40	0.209	0.151	37.07	-118.80	37.7	10200
11221700[P]	101	38	-0.472	0.159	36.76	-119.16	127	2636
11222099[A]	102	104	0.533	0.084	36.91	-118.87	1681	7213
11224500	103	60	-0.505	0.133	36.26	-120.57	95.8	2639
11226500[R]	104	44	0.151	0.122	37.58	-119.11	249	9113
11228500	105	42	0.640	0.241	37.59	-119.32	47.8	8890
11230500[P]	106	82	0.589	0.113	37.34	-118.86	52.5	10636
11237500[P]	107	79	0.493	0.105	37.18	-119.16	22.9	7997
11242400[P]	108	41	0.116	0.155	37.44	-119.53	16.9	6628
11251099[A]	109	92	0.635	0.108	37.33	-119.22	1678	6903
11253310	110	49	-0.643	0.450	36.39	-120.54	46.4	2542
11257500	111	75	-0.370	0.088	37.36	-119.68	133	3201
11259099[A]	112	69	-0.898	0.173	37.39	-119.86	303	1769
11259999[A]	113	49	-0.840	0.295	37.93	-120.77	208	572
11264500[R]	114	91	0.428	0.085	37.71	-119.42	181	9009
11266500[R]	115	90	0.627	0.109	37.74	-119.49	321	8459
11268500[R]	116	44	0.467	0.184	37.67	-119.70	911	5988
11270099[A]	117	97	0.140	0.064	37.67	-119.76	1038	5473
11274500	118	75	-1.015	0.261	37.29	-121.32	134	1551
11274630[P]	119	48	-0.861	0.237	37.44	-121.36	72.8	1835
11281000[P]	120	80	-0.081	0.078	37.81	-119.83	87.0	5549

Note: A = U.S. Army Corp of Engineers Gage Site, E = Eastern Sierra Lahontan Region Gage Site (removed), L = Gage site with lack of LP3 fit (removed), P = Pristine Gage Site, R = Redundant Gage Site (removed)

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	Years of Record	EMA At-Site Log Skew	EMA At-Site Var(Log Skew)	Basin Centroid Location Lat	Basin Centroid Location Lon	Drainage Area (mi^2)	Mean Basin Elevation (ft)
11282000[P]	121	86	0.280	0.079	37.87	-119.78	73.5	6203
11283500[P]	122	33	0.137	0.145	38.07	-120.00	144	5890
11284400	123	37	-0.898	0.296	37.76	-120.02	16.1	3151
11286500	124	44	-0.219	0.183	37.97	-120.35	97.2	2253
11288099[A]	125	104	0.452	0.069	37.96	-119.87	1532	5889
11289950[P]	126	35	-0.239	0.193	37.73	-120.63	193	430
11292500[P]	127	44	0.298	0.157	38.41	-119.76	67.5	8176
11294500[P]	128	67	0.432	0.108	38.41	-120.04	163	6855
11299599[A]	129	68	-0.003	0.088	38.26	-120.08	904	5662
11308999[A]	130	35	-0.972	0.360	38.20	-120.59	372	1991
11312000	131	54	-0.395	0.467	38.17	-120.99	47.6	271
11315000[P]	132	77	0.333	0.089	38.56	-120.14	21.0	7414
11316800[P]	133	46	-0.151	0.140	38.42	-120.34	20.8	4656
11317000[P]	134	95	-0.412	0.079	38.41	-120.36	68.4	4385
11318500[P]	135	73	-0.390	0.101	38.35	-120.38	75.1	4099
11323599[A]	136	93	0.526	0.091	38.45	-120.32	628	4918
11329500	137	50	-0.685	0.181	38.37	-120.85	324	1205
11335099[A]	138	91	-0.687	0.104	38.60	-120.63	535	3064
11342000	139	62	0.129	0.101	41.16	-122.37	425	4147
11348500	140	77	-0.317	0.095	41.41	-120.52	1431	5341
11370599[A,R]	141	73	-0.681	0.127	41.12	-121.25	7368	4699
11371000	142	44	0.508	0.338	40.82	-122.60	115	3244
11374000	143	57	-0.658	0.149	40.66	-121.99	425	2251
11376000[P]	144	66	-0.666	0.143	40.31	-122.70	927	2221
11376550	145	44	-0.083	0.162	40.45	-121.75	357	4074
11379000	146	41	-0.857	0.270	40.25	-121.78	123	3356
11381500	147	78	-0.155	0.082	40.23	-121.66	131	3961
11382000	148	76	-0.312	0.091	39.91	-122.78	203	4146
11383500	149	90	-0.379	0.078	40.18	-121.57	208	4199
11384000	150	57	-0.474	0.154	39.99	-121.65	72.4	3095

Note: A = U.S. Army Corp of Engineers Gage Site, E = Eastern Sierra Lahontan Region Gage Site (removed), L = Gage site with lack of LP3 fit (removed), P = Pristine Gage Site, R = Redundant Gage Site (removed)

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	Years of Record	EMA At-Site Log Skew	EMA At-Site Var(Log Skew)	Basin Centroid Location Lat	Basin Centroid Location Lon	Drainage Area (mi^2)	Mean Basin Elevation (ft)
11388099[A]	151	35	-0.638	0.260	39.57	-122.62	742	2411
11390000[P]	152	76	-0.478	0.097	39.97	-121.56	147	3716
11402000	153	73	-0.680	0.116	39.94	-120.93	184	4813
11407099[A,R]	154	96	-0.217	0.066	36.92	-118.84	3624	7634
11407500	155	37	0.025	0.219	39.43	-121.34	30.6	1642
11413000[R]	156	76	-0.090	0.075	39.59	-120.70	250	5681
11413599[A]	157	55	-0.324	0.129	39.59	-120.85	486	4912
11414000[P]	158	53	0.343	0.110	39.33	-120.44	51.8	6891
11427000	159	65	-0.371	0.112	39.17	-120.67	341	4358
11427700[P]	160	46	0.240	0.145	39.16	-120.46	9.94	6398
11428000[L]	161	32	-0.375	0.237	38.95	-120.20	31.4	7674
11431800[P]	162	46	0.031	0.136	38.91	-120.49	11.7	5146
11439500[P]	163	84	0.089	0.070	38.74	-120.15	193	7092
11446599[A]	164	94	-0.035	0.054	38.94	-120.59	1888	4277
11449500	165	60	-0.935	0.193	38.86	-122.79	36.6	2645
11451100[R]	166	35	-0.553	0.241	39.20	-122.73	60.2	2989
11451500	167	51	-1.112	0.297	39.14	-122.64	197	2453
11453500	168	57	-1.045	0.176	38.77	-122.62	113	1824
11456000	169	66	-1.224	0.203	38.56	-122.54	78.8	1024
11461000	170	56	-0.272	0.101	39.29	-123.26	100	1468
11464500	171	39	-0.995	0.247	38.83	-123.15	87.8	1193
11468000	172	56	-0.103	0.154	39.06	-123.42	303	1126
11468500	173	55	0.070	0.113	39.43	-123.57	106	918
11469000	174	58	-0.865	0.192	40.19	-124.08	245	1354
11472200	175	39	0.160	0.133	39.50	-123.39	161	1961
11473900[R]	176	41	0.138	0.282	39.79	-123.07	745	3685
11475500[P]	177	30	-0.076	0.144	39.72	-123.65	43.9	1534
11475560	178	39	0.035	0.172	39.71	-123.60	6.50	2792
11475800[R]	179	38	-0.017	0.181	39.76	-123.61	248	2053
11476500	180	66	-0.317	0.093	39.90	-123.70	537	1726

Note: A = U.S. Army Corp of Engineers Gage Site, E = Eastern Sierra Lahontan Region Gage Site (removed), L = Gage site with lack of LP3 fit (removed), P = Pristine Gage Site, R = Redundant Gage Site (removed)

Table A (Continued)

USGS Hydrologic Unit Code	Site Index #	Years of Record	EMA At-Site		Basin Centroid Location		Drainage Area (mi^2)	Mean Basin Elevation (ft)
			Log Skew	Var(Log Skew)	Lat	Lon		
11477000	181	95	-0.166	0.066	39.86	-123.37	3113	2577
11478500	182	67	-0.513	0.135	40.42	-123.63	222	3032
11481200	183	51	-0.373	0.212	41.02	-124.00	40.5	1096
11481500	184	35	-0.285	0.298	40.81	-123.76	67.7	2935
11482500	185	56	-0.797	0.152	41.06	-123.89	277	1831
11519500	186	65	0.098	0.164	41.48	-122.84	653	4333
11521500	187	54	0.135	0.149	41.90	-123.48	120	3696
11522500	188	82	-0.302	0.080	41.29	-123.19	751	4261
11523200	189	49	0.315	0.251	41.23	-122.64	149	5340
11528700[R]	190	40	-0.636	0.217	40.47	-123.23	765	3683
11530000[P]	191	33	-0.321	0.187	40.79	-123.12	2854	3770
11532500	192	75	-0.595	0.119	41.83	-123.87	614	2521

Note: A = U.S. Army Corp of Engineers Gage Site, E = Eastern Sierra Lahontan Region Gage Site (removed), L = Gage site with lack of LP3 fit (removed), P = Pristine Gage Site, R = Redundant Gage Site (removed)

APPENDIX B

REMOVED GAUGE SITES FROM THE EASTERN SIERRA AND LAHONTAN DESERT REGIONS

This appendix contains Table B, which lists the 17 gauge sites removed from the study in the Eastern Sierra Lahontan Desert Region. See Section 3.2.1 for a discussion of why these sites are removed from the California regional skew study.

Table B: 17 gauge sites located in the Eastern Sierra and Lahontan Desert Regions removed for the California regional skew study.

Index #	USGS #	SITE NAME
1	9423350	CARUTHERS C NR IVANPAH CA
2	10251300	AMARGOSA RIVER AT TECOPA, CA
3	10257600	MISSION C NR DESERT HOT SPRINGS CA
4	10258000	TAHQUITZ C NR PALM SPRINGS CA
5	10259000	ANDREAS C NR PALM SPRINGS CA
6	10260500	DEEP C NR HESPERIA CA
7	10261000	WF MOJAVE R NR HESPERIA CA
8	10263500	BIG ROCK C NR VALYERMO CA
9	10263900	BUCKHORN C NR VALYERMO CA
10	10264000	LITTLE ROCK C AB LTTLE ROCK RES NR LITTLEROCK CA
11	10264600	OAK C NR MOJAVE CA
12	10265200	CONVICT C NR MAMMOTH LAKES CA
13	10265700	ROCK C A LITTLE ROUND VALLEY NR BISHOP CA
14	10267000	PINE C A DIVISION BOX NR BISHOP CA
15	10268700	SILVER CYN C NR LAWS CA
16	10281800	INDEPENDENCE C BL PINYON C NR INDEPENDENCE CA
17	10286000	COTTONWOOD C NR OLANCHA CA

APPENDIX C

REMOVED REDUNDANT GAUGE SITES

This appendix contains Table C which lists the 16 gauge sites removed from the California regional skew study due to redundancy. See Section 3.2.2 for a detailed explanation of redundant sites.

Table C: 16 redundant gauge sites removed from the California regional skew study

Index #	USGS #	SITE NAME
19	10295500	L WALKER R NR BRIDGEPORT, CA
60	11138500	SISQUOC R NR SISQUOC CA
75	11156500	SAN BENITO R NR WILLOW CREEK SCHOOL CA
95	11209500	NF KAWEAH R A KAWEAH CA
99	11213500	KINGS R AB NF NR TRIMMER CA
104	11226500	SAN JOAQUIN R A MILLER CROSSING CA
114	11264500	MERCED R A HAPPY ISLES BRIDGE NR YOSEMITE CA
115	11266500	MERCED R A POHONO BRIDGE NR YOSEMITE CA
116	11268500	MERCED R A BAGBY CA
141	11370599[A]	SACRAMENTO R A KESWICK CA
154	11407099[A]	FEATHER R A OROVILLE CA
156	11413000	N YUBA R BL GOODYEARS BAR CA
166	11451100	NF CACHE C A HOUGH SPRING NR CLEARLAKE OAKS CA
176	11473900	MF EEL R NR DOS RIOS CA
179	11475800	SF EEL R A LEGGETT CA
190	11528700	SF TRINITY R BL HYAMPOM CA

Note: A indicates an U.S. Army Corp of Engineers gauge site

APPENDIX D

GAUGE SITES USED TO DEVELOP CROSS-CORRELATION MODEL

This appendix contains Table D which lists the 21 gauge sites used to develop the cross-correlation of peak annual flows model for the California regional skew study. See Section 3.2.3 for a detailed discussion of the cross-correlation of peak annual flows and how it is used in the regional skew analysis.

Table D: 21 gauge sites used to develop cross-correlation of annual peak flows for the California regional skew study

Index #	USGS #	SITE NAME
31	11055500	PLUNGE C NR EAST HIGHLANDS CA
32	11055800	CITY C NR HIGHLAND CA
33	11058500	E TWIN C NR ARROWHEAD SPRINGS CA
34	11058600	WATERMAN CANYON CREEK NR ARROWHEAD SPRINGS CA
40	11098000	ARROYO SECO NR PASADENA CA
55	11132500	SALSIPUEDES C NR LOMPOC CA
72	11152000	ARROYO SECO NR SOLEDAD CA
81	11160500	SAN LORENZO R A BIG TREES CA
86	11169500	SARATOGA C A SARATOGA CA
91	11189500	SF KERN R NR ONYX CA
106	11230500	BEAR C NR LAKE THOMAS A EDISON CA
107	11237500	PITMAN C BL TAMARACK C CA
120	11281000	SF TUOLUMNE R NR OAKLAND RECREATION CAMP CA
121	11282000	M TUOLUMNE R A OAKLAND RECREATION CAMP CA
128	11294500	NF STANISLAUS R NR AVERY CA
132	11315000	COLE C NR SALT SPRINGS DAM CA
134	11317000	MF MOKELUMNE R A WEST POINT CA
144	11376000	COTTONWOOD C NR COTTONWOOD CA
152	11390000	BUTTE C NR CHICO CA
163	11439500	SF AMERICAN R NR KYBURZ(RIVER ONLY) CA
191	11530000	TRINITY R A HOOPA CA

APPENDIX E

PRISTINE GAUGE SITES USED IN EXTENDED BAYESIAN GLS ANALYSIS

This appendix contains Table E which lists the 64 gauge sites in the pristine data set. The pristine data set is a subset of the larger data set used in the WLS estimation of the regression parameters, and it does not include sites with historical information, zero flows, or low outliers as determined by EMA, or any reconstructed flow records. See Section 3.3.2 for a detailed description of the pristine data set and its use in the California regional skew analysis.

Table E: Pristine gauge sites (64 sites) used in modified Bayesian GLS analysis

Index #	USGS #	SITE NAME
18	10291500	BUCKEYE CREEK NEAR BRIDGEPORT, CA
20	10296500	W WALKER R NR COLEVILLE, CA
21	10336610	UPPER TRUCKEE RIVER AT SOUTH LAKE TAHOE CALIF
22	10336676	WARD C AT HWY 89 NR TAHOE PINES CA
23	10336780	TROUT CREEK NR TAHOE VALLEY CALIF
24	10343500	SAGEHEN C NR TRUCKEE CA
25	10358500	WILLOW C NR SUSANVILLE CA
30	11046100	LAS FLORES C NR OCEANSIDE CA
31	11055500	PLUNGE C NR EAST HIGHLANDS CA
32	11055800	CITY C NR HIGHLAND CA
33	11058500	E TWIN C NR ARROWHEAD SPRINGS CA
34	11058600	WATERMAN CANYON CREEK NR ARROWHEAD SPRINGS CA
35	11063000	CAJON C NR KEENBROOK CA
36	11073470	CUCAMONGA C NR UPLAND CA
37	11075800	SANTIAGO C A MODJESKA CA
38	11084000	ROGERS C NR AZUSA CA
40	11098000	ARROYO SECO NR PASADENA CA
41	11100000	SANTA ANITA C NR SIERRA MADRE CA
42	11100500	LITTLE SANTA ANITA C NR SIERRA MADRE CA
44	11105850	ARROYO SIMI NR SIMI CA
45	11107745	SANTA CLARA R AB RR STATION NR LANG CA
55	11132500	SALSIPUEDES C NR LOMPOC CA
56	11134800	MIGUELITO C A LOMPOC CA
57	11136100	SAN ANTONIO C NR CASMALIA CA
59	11138000	HUASNA R NR SANTA MARIA CA
61	11139500	TEPUSQUET C NR SISQUOC CA
66	11143500	SALINAS R NR POZO CA
71	11151300	SAN LORENZO C BL BITTERWATER C NR KING CITY CA
72	11152000	ARROYO SECO NR SOLEDAD CA
73	11152540	EL TORO C NR SPRECKELS CA
79	11159200	CORRALITOS C A FREEDOM CA
81	11160500	SAN LORENZO R A BIG TREES CA
86	11169500	SARATOGA C A SARATOGA CA
90	11182500	SAN RAMON C A SAN RAMON CA
91	11189500	SF KERN R NR ONYX CA
92	11200800	DEER C NR FOUNTAIN SPRINGS CA
93	11203500	TULE R NR PORTERVILLE CA
94	11204500	SF TULE R NR SUCCESS CA
96	11210500	KAWEAH R NR THREE RIVERS CA
97	11211300	DRY C NR LEMONCOVE CA

Table E (continued)

Index #	USGS #	SITE NAME
100	11214000	NF KINGS R BL MEADOWBROOK CA
101	11221700	MILL C NR PIEDRA CA
106	11230500	BEAR C NR LAKE THOMAS A EDISON CA
107	11237500	PITMAN C BL TAMARACK C CA
108	11242400	NF WILLOW C NR SUGAR PINE CA
119	11274630	DEL PUERTO C NR PATTERSON CA
120	11281000	SF TUOLUMNE R NR OAKLAND RECREATION CAMP CA
121	11282000	M TUOLUMNE R A OAKLAND RECREATION CAMP CA
122	11283500	CLAVEY R NR BUCK MEADOWS CA
126	11289950	DRY C NR MODESTO CA
127	11292500	CLARK FORK STANISLAUS R NR DARDANELLE CA
128	11294500	NF STANISLAUS R NR AVERY CA
132	11315000	COLE C NR SALT SPRINGS DAM CA
133	11316800	FOREST C NR WILSEYVILLE CA
134	11317000	MF MOKELUMNE R A WEST POINT CA
135	11318500	SF MOKELUMNE R NR WEST POINT CA
144	11376000	COTTONWOOD C NR COTTONWOOD CA
152	11390000	BUTTE C NR CHICO CA
158	11414000	S YUBA R NR CISCO CA
160	11427700	DUNCAN CYN C NR FRENCH MEADOWS CA
162	11431800	PILOT C AB STUMPY MEADOWS RES CA
163	11439500	SF AMERICAN R NR KYBURZ(RIVER ONLY) CA
177	11475500	SF EEL R NR BRANSCOMB CA
191	11530000	TRINITY R A HOOPA CA

REFERENCES

Cohn, T.A., W.L. Lane, and W.G. Baier (1997), An algorithm, for computing moments-based flood quantile estimates when historical flood information is available, *Water Resour. Res.,* 33(9), 2089-2096.

Feaster, T.D., Gotvald, A.J., and Weaver, J.C., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume *3*, South Carolina: U.S. Geological Survey

FRAP. (2000), California Dept. of Forestry and Fire Protection, Precipitation Zones: Mean Annual 1900-1960, http://frap.cdf.ca.gov/data/frapgismaps/download.asp.

FRAP. (2003), California Dept. of Forestry and Fire Protection, Land Cover, http://frap.cdf.ca.gov/data/frapgismaps/download.asp.

Gotvald, A.J., Feaster, T.D., and Weaver, J.C., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 1, Georgia: U.S. Geological Survey Scientific Investigations Report 2009–5043, 120 p.

Griffis, V. W., and J. R. Stedinger, (2007), The Use of GLS Regression in Regional Hydrologic Analyses, *J. of Hydrology,* 344(1-2), 82-95, [doi:10.1016/j.jhydrol.2007.06.023].

Griffis, V.W., and J. R. Stedinger, (2009), The Log-Pearson Type 3 Distribution and its Application in Flood Frequency Analysis, 3. Sample Skew and Weighted Skew Estimators, *J. of Hydrol. Engineering* 14(2), 121-130.

Gruber, Andrea M., Dirceu S. Reis Jr., and Jery R. Stedinger, (2007), Models of Regional Skew Based on Bayesian GLS Regression, Paper 40927-3285, World Environmental & Water Resources Conference - Restoring our Natural Habitat, K.C. Kabbes editor, Tampa, Florida, May 15-18.

Gruber, A.M., J.R. Stedinger, (2008), Models of LP3 Regional Skew, Data Selection, and Bayesian GLS Regression, World Environmental & Water Resources Conference 2008 AHUPUA‘A, paper 596, R. Babcock and R. Walton (eds.), Amer. Soc. of Civil Engineers, Hawaii, May 12-16.

Hardison, C.H., (1971), Prediction error of regression estimates of streamflow characteristics at ungaged sites, U.S. Geological Survey Professional Paper 750-C, C228-C236.

Interagency Advisory Committee on Water Data (1982), Guidelines for Determining Flood Flow Frequency, Bulletin #17B, U.S. Department of the Interior, U.S. Geological Survey, Office of Water Data Coordination, Reston Virginia.

Martins, E.S., and J.R. Stedinger, (2002), Cross-correlation among estimators of shape, *Water Resources Research,* 38(11), doi: 10.1029/2002WR001589, 26 November.

Mount, J.F., (1995), California rivers and streams: the conflict between fluvial process and land use, University of California Press, 359pp.

Reis, D. S., Jr., J. R. Stedinger, and E. S. Martins (2005), Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation, Water Resour. Res., 41, W10419, doi:10.1029/2004WR003445.

Stedinger, J.R., and V.W. Griffis, (2008), Flood Frequency Analysis in the United States: Time to Update (editorial), *J. of Hydrology, 13*(4), 199-204, April 2008.

Stedinger, J.R., and G.D. Tasker, (1985), Regional Hydrologic Analysis, 1. Ordinary, Weighted and Generalized Least Squares Compared, *Water Resources Research, 21*(9), 1421-1432. [with correction, *Water Resour. Res. 22*(5), 844, 1986.]

Tasker, G.D., and J.R. Stedinger, (1986), Estimating Generalized Skew With Weighted Least Squares Regression, *Journal of Water Resources Planning and Management, 112*(2), 225-237.

Veilleux, A. G., (2009). Bayesian GLS Regression for Regionalization of Hydrologic Statistics, Floods and Bullentin 17 Skew, M.S. Thesis, Cornell University, August.

Weaver, J.C., Feaster, T.D., and Gotvald, A.J., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 2, North Carolina: U.S. Geological Survey Scientific Investigations Report, 2009-5158.

CHAPTER 4

BAYESIAN WLS/GLS REGRESSION FOR REGIONAL SKEWNESS ANALYSIS FOR REGIONS WITH LARGE CROSS-CORRELATIONS AMONG FLOOD FLOWS

4.1 Introduction

Methodological advances in regional log-space skewness analyses to support flood frequency analysis with the LP3 distribution are summarized in this chapter. Parrett *et al*. [2011] report a regional analysis of skewness coefficients (log-space skewness for the LP3 distribution) for the non-desert regions of California. They found that the cross-correlations between annual peak discharges in California are often large relative to other studies. The large cross-correlations present difficulties in the regional skewness analysis because a Bayesian Generalized Least Squares (B-GLS) analysis seeks to exploit the cross-correlations among the sample skewness estimates to obtain the best possible estimates of the model parameters. If the cross-correlations are large, the Generalized Least Squares (GLS) estimators can become relatively complicated as a result of the effort to find the most efficient estimator of the parameters. Unfortunately, the precision of the cross-correlation estimates between any two particular sites is not of sufficient precision to justify the sophisticated weights (both positive and negative) that the B-GLS analysis generates. Thus, an alternate regression procedure using both Weighted Least Squares (WLS) and GLS is developed so that the regional skewness analysis can provide both stable and defensible results.

A second California skew study considered rainfall flood volumes corresponding to 1 to 30 day durations. These rainfall flood volumes exhibited even larger cross-correlations than did

concurrent annual maxima. For that analysis a new WLS/GLS regression framework was developed. It uses an Ordinary Least Squares (OLS) analysis to fit an initial regional skewness model that is used to generate a stable regional skewness coefficient estimate for each site. That estimate is the basis for computing the variance of each at-site skewness estimator employed in the WLS analysis. Then, WLS is used to generate an estimator of the regional skewness model parameters. Finally, B-GLS is used to estimate the precision of that WLS parameter estimator, and to estimate the model error variance and the precision of that variance estimator. The steps of this alternative procedure are described in detail below. Regional skewness results using a dataset from the Southeastern U.S. illustrate application of the methodology.

4.2 Bayesian WLS/GLS Regression Framework

4.2.1 OLS Analysis

The first step in the regional skewness analysis is the estimation of a regional skewness model using OLS. This is an iterative procedure. For the first iteration, the constant model is considered. After the subsequent WLS and GLS (Sections 4.2.2 and 4.2.3 below) analysis are performed to determine which basin characteristics are statistically significant in explaining regional skewness, the OLS regional model can be expanded to incorporate those additional basin characteristics, and thus to better describe variations in skewness from watershed-to-watershed.

The at-site skewness estimates are unbiased by using the correction factor developed by Tasker and Stedinger [1986] and employed in Reis *et al*. [2005]. The unbiased at-site skewness estimator is

$$\hat{\gamma}_i = \left[1 + \frac{6}{N_i}\right] G_i \tag{4.1}$$

Here $\hat{\gamma}_i$ is the unbiased at-site sample skewness estimate for site *i*, N_i is the systematic record length at site i, G_i is the traditional biased at-site skewness estimator for site *i* or the EMA skewness estimate if the site has zero flows, low outliers or historical peaks. When unbiasing the skew, N_i is the number of systematic peaks and thus, additional information provided by any historical flood period is neglected.

The regional regression parameters estimated by OLS, $\hat{\boldsymbol{\beta}}_{OLS}$, are calculated as

$$\hat{\boldsymbol{\beta}}_{OLS} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \hat{\boldsymbol{\gamma}} \tag{4.2}$$

where the superscript *T* denotes a matrix transpose, $\mathbf{X}$ is (*n* x *k*) matrix of basin characteristics, $\hat{\boldsymbol{\gamma}}$ is an (*n* x 1) vector of the unbiased at-site sample skewness estimates, *n* is the number of gauge sites, and *k* is the number of basin parameters including a column of ones to estimate the constant. After computing $\hat{\boldsymbol{\beta}}_{OLS}$, the unbiased and relatively stable regional vector-estimate of the skewness for all sites is

$$\tilde{\mathbf{y}}_{OLS} = \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS} \tag{4.3}$$

These estimated regional skewness values $\tilde{\mathbf{y}}_{OLS}$ are then used in conjunction with the at-site record lengths to estimate the variance of the at-site sample skewness. The variance of the unbiased at-site skewness includes the correction factor developed by Tasker and Stedinger [1986]:

$$Var\left[\hat{\gamma}_i\right] = \left[1 + \frac{6}{N_i}\right]^2 Var\left[G_i\right] \tag{4.4}$$

The unbiased at-site regional skewness variances in Equation 4.4 are calculated using the equations developed by Griffis and Stedinger [2009]. These at-site variances of the skewness are based on the regional OLS estimator of the skewness coefficient instead of the at-site skewness estimator, thus making the weights in the subsequent steps relatively independent of the at-site skewness estimates. The computation generally neglects complicating factors such as zero flow years, censored observations/ low outliers, and modest historical records.

4.2.2 WLS Analysis

A Weighted Least Squares analysis is used to develop estimators of the regression coefficients for each regional skewness model. The WLS analysis explicitly reflects variations in record length, but neglects cross correlations thereby avoiding the problems experienced with GLS parameter estimators. After the regression model coefficients are determined with WLS, the precision of a model and the precision of the estimated regression coefficients are estimated using an appropriate GLS analysis (Section 4.2.3).

The first step in the WLS analysis is to use Bayesian-WLS (B-WLS) to estimate the model error variance, $\sigma^2_{\delta,B-WLS}$ [Reis *et al*., 2005]. Using a B-WLS approach to estimate the model error variance, avoids the possible pitfall of estimating the model error variance as zero, which can occur when using Method-of-Moments WLS. It is important to note that the Bayesian analysis produces an estimate of the distribution of the model error variance, however only the mean model error variance estimator, $\sigma^2_{\delta,B-WLS}$, is considered in this analysis. Given the model error variance estimator $\sigma^2_{\delta,B-WLS}$, a WLS analysis is used to generate the weight matrix, **W**, needed to compute estimates of the regression parameters $\hat{\boldsymbol{\beta}}_{WLS}$. In order to compute **W**, a

diagonal covariance matrix, $\mathbf{\Lambda_{WLS}}\left(\sigma^2_{\delta,B-WLS}\right)$, is created. As specified in Equation (4.5), the diagonal elements of the covariance matrix are the sum of the estimated model error variance, $\sigma^2_{\delta,B-WLS}$, and the variance of the unbiased at-site skewness estimator, $Var\left[\hat{\gamma}_i\right]$, which depends upon on the at-site record length, and the estimate of the regional skewness for each site calculated by OLS, $\tilde{\mathbf{y}}_{OLS}$. The off-diagonal elements of $\mathbf{\Lambda_{WLS}}\left(\sigma^2_{\delta,B-WLS}\right)$ are zero, because cross-correlations between gage sites are not considered in a WLS analysis. Thus the (n x n) covariance matrix, $\mathbf{\Lambda_{WLS}}\left(\sigma^2_{\delta,B-WLS}\right)$ is given by

$$\mathbf{\Lambda_{WLS}}\left(\sigma^2_{\delta,B-WLS}\right) = \sigma^2_{\delta,B-WLS}\mathbf{I} + diag\left(Var\left[\hat{\boldsymbol{\gamma}}\right]\right) \tag{4.5}$$

where, **I** is an (n x n) identity matrix, n is the number of gage sites in the study, and $diag\left(Var\left[\hat{\boldsymbol{\gamma}}\right]\right)$ is an (n x n) matrix containing the variance of the unbiased at-site sample skewness estimators, $Var\left[\hat{\gamma}_i\right]$, on the diagonal and zeros on the off-diagonal. Using that covariance matrix, the WLS weights are calculated as

$$\mathbf{W} = \left[\mathbf{X}^T\mathbf{\Lambda_{WLS}}\left(\sigma^2_{\delta,B\text{-}WLS}\right)^{-1}\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{\Lambda_{WLS}}\left(\sigma^2_{\delta,B-WLS}\right)^{-1} \tag{4.6}$$

where **W** is (k x n) the matrix of weights, **X** is the (n x k) matrix of basin parameters, and k is the number of columns in the **X** matrix. These weights are used to compute the final estimates of the regression parameters $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}}_{WLS} = \mathbf{W}\hat{\boldsymbol{\gamma}} \tag{4.7}$$

where $\hat{\boldsymbol{\beta}}_{WLS}$ is the (k x 1) vector of estimated regression parameters.

4.2.3 Bayesian GLS Analysis

After the regression model coefficients, $\hat{\boldsymbol{\beta}}_{WLS}$ **,** and weights, **W**, are determined with a WLS analysis (Section 4.2.2), the precision of the fitted model and the precision of the regression coefficients are estimated using a Bayesian-GLS (B-GLS) analysis. Using the B-GLS regression framework for regional skew developed by Reis *et al*. [2005], the posterior probability density function for $\sigma^2_{\delta,B-GLS}$ is

$$f\left(\sigma^2_{\delta,B-GLS} \mid \hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}_{WLS}\right) \propto \xi\left(\sigma^2_{\delta,B-GLS}\right) * \left|\boldsymbol{\Lambda}_{\mathbf{GLS}}\left(\sigma^2_{\delta,B-GLS}\right)\right|^{-0.5} * \exp\left[-0.5\left(\hat{\boldsymbol{\gamma}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{WLS}\right)^T \left(\boldsymbol{\Lambda}_{\mathbf{GLS}}\left(\sigma^2_{\delta,B-GLS}\right)\right)^{-1} \left(\hat{\boldsymbol{\gamma}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{WLS}\right)\right] \tag{4.8}$$

where $\hat{\boldsymbol{\gamma}}$ represents the skew data and $\xi\left(\sigma^2_{\delta,B-GLS}\right)$ is the exponential prior for the model error variance described by

$$\xi\left(\sigma^2_{\delta,B-GLS}\right) = \lambda e^{-\lambda\left(\sigma^2_{\delta,B-GLS}\right)}, \quad \sigma^2_{\delta,B-GLS} > 0 \tag{4.9}$$

A value for lambda of 10 was adopted, corresponding to a mean model error variance of 1/10. That prior assigns a 63% probability to the interval [0, 0.1], 86% probability to the interval [0,0.2], and 95% probability to the interval [0, 0.3].

The mean B-GLS model error variance, $\sigma^2_{\delta,B-GLS}$, can then be used to compute the precision of the regression parameters $\hat{\boldsymbol{\beta}}_{WLS}$ that were calculated with the WLS weights **W**. The GLS covariance matrix for the WLS **β**-estimator, $\hat{\boldsymbol{\beta}}_{WLS}$, is simply

$$\boldsymbol{\Sigma}\left(\hat{\boldsymbol{\beta}}_{WLS}\right) = \mathbf{W}\boldsymbol{\Lambda}_{\mathbf{GLS}}\left(\sigma^2_{\delta,B-GLS}\right)\mathbf{W}^T \tag{4.10}$$

where $\boldsymbol{\Lambda}_{\mathbf{GLS}}\left(\sigma^2_{\delta,B-GLS}\right)$ is an (*n* x *n*) GLS covariance matrix calculated as

$$\boldsymbol{\Lambda}_{\mathbf{GLS}}\left(\sigma^2_{\delta,B-GLS}\right) = \sigma^2_{\delta,B-GLS}\mathbf{I} + \boldsymbol{\Sigma}\left(\hat{\gamma}\right) \tag{4.11}$$

Here $\mathbf{I}$ is an (n x n) identity matrix, $\boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}})$ is a full (n x n) matrix containing the sampling variances of the unbiased skewness estimators, $Var[\hat{\gamma}_i]$, and the covariances of the skewness estimators $\hat{\gamma}_i$. The off-diagonal values of $\boldsymbol{\Sigma}(\hat{\boldsymbol{\gamma}})$ are determined by the the cross-correlation of concurrent systematic annual peak flows and the *cf* factor [Martins and Stedinger, 2002, eqn. 3]. When calculating the *cf* factor using the ratio between the number of concurrent peak flows at a pair of sites and the total number of peak flows at both sites, only the systematic records are considered. Thus, any additional information provided by a historical flood period included in the EMA analysis would have been neglected in the calculation of the cross-correlation of peak flows and the *cf* factor.

4.3 Diagnostic Statistics for WLS/GLS Regional Analysis

This section discusses several diagnostic statistics that are useful for describing the precision of model predictions, and investigating whether particular sites have unusual leverage or influence upon the results. The variance of prediction is a common metric used to choose which of several models provides the most accurate estimator of the y-variable, because it combines both the model error variance and the sampling error in the model parameters.

4.3.1 Variance of Prediction

The variance of prediction depends upon whether one is considering a new site that was not used to derive the estimate of the parameters (see Equation 4.11), or an old site whose sample estimator of the skewness was used to compute the estimates of the parameters . For an

old site, there is correlation between the error in the at-site estimator and the estimated parameters.

The Bayesian variance of prediction of the skewness at a new site *i* with basin characteristics $\mathbf{x}_i$ is given by

$$\begin{aligned} VP_{new}(i) &= \underset{\sigma^2_{\delta,B-GLS}}{E}\left[\sigma^2_{\delta,B-GLS} + \mathbf{x}_i\mathbf{W}\left(\mathbf{\Lambda_{GLS}}\left(\sigma^2_{\delta,B-GLS}\right)\right)\mathbf{W}^T\mathbf{x}_i^T\right] \\ &= \underset{\sigma^2_{\delta,B-GLS}}{E}\left\{\sigma^2_{\delta,B-GLS}\right\} + \mathbf{x}_i Var\left[\hat{\boldsymbol{\beta}}_{WLS}\right]\mathbf{x}_i^T \end{aligned} \tag{4.12}$$

wherein $\sigma^2_{\delta,B-GLS}$ reflects the underlying error in the model, and $\mathbf{x}_i\mathbf{W}\Lambda\mathbf{W}^T\mathbf{x}_i^T$ reflects the precision with which the model parameters are estimated and the possible errors that would occur in predicting the skewness at a site with basin characteristics $\mathbf{x}_i$. However, if the predictions are made for those *n* old sites used in the analysis, the Bayesian variance of prediction is given by

$$\begin{aligned} VP_{old}(i) &= \underset{\sigma^2_{\delta,B-GLS}}{E}\left[\sigma^2_{\delta,B-GLS} + \mathbf{x}_i\mathbf{W}\left(\mathbf{\Lambda_{GLS}}\left(\sigma^2_{\delta,B-GLS}\right)\right)\mathbf{W}^T\mathbf{x}_i^T - 2*\sigma^2_{\delta,B-GLS}*\mathbf{x}_i\mathbf{W}e_i\right] \\ &= \underset{\sigma^2_{\delta,B-GLS}}{E}\left\{\sigma^2_{\delta,B-GLS}\right\} + \mathbf{x}_i Var\left[\hat{\boldsymbol{\beta}}_{WLS}\right]\mathbf{x}_i^T - \underset{\sigma^2_{\delta,B-GLS}}{E}\left\{\sigma^2_{\delta,B-GLS}\right\}\left(2\mathbf{x}_i\mathbf{W}e_i\right) \end{aligned} \tag{4.13}$$

wherein e_i is an (*n* x 1) column vector with one at the i^{th} row and zero otherwise.

4.3.2 Leverage

The leverage measure, $\mathbf{H}^*$, for a GLS regression as described by Tasker and Stedinger [1989, eqn. 23] is

$$\mathbf{H}^* = \mathbf{X}\left\{\mathbf{X}^T\mathbf{\Lambda}^{-1}_{GLS}\left(\sigma^2_{\delta,MM-GLS}\right)\mathbf{X}\right\}^{-1}\mathbf{X}^T\mathbf{\Lambda}^{-1}_{GLS}\left(\sigma^2_{\delta,MM-GLS}\right) \tag{4.14}$$

where $\mathbf{\Lambda}^{-1}_{GLS}\left(\sigma^2_{\delta,MM-GLS}\right)$ is the inverse of the (*n* x *n*) covariance matrix $\mathbf{\Lambda}_{GLS}\left(\sigma^2_{\delta,MM-GLS}\right)$. With the WLS/GLS methodology used in this study, the WLS step selects weights, **W**, used to

estimate the coefficients, and thus determines the leverage that should be associated with each observation. In calculating the leverage, a diagonal covariance matrix is used with the B-WLS model error variance. Thus, using the framework for leverage provided by Tasker and Stedinger [1989], the leverage for this study is

$$\begin{aligned}\mathbf{H}^*_{WLS} &= \mathbf{XW} \\ &= \mathbf{X}\left\{\mathbf{X}^T\boldsymbol{\Lambda}^{-1}_{WLS}\left(\sigma^2_{\delta,B-WLS}\right)\mathbf{X}\right\}^{-1}\mathbf{X}^T\boldsymbol{\Lambda}^{-1}_{WLS}\left(\sigma^2_{\delta,B-WLS}\right)\end{aligned} \tag{4.15}$$

where $\boldsymbol{\Lambda}^{-1}_{WLS}\left(\sigma^2_{\delta,B-WLS}\right)$ is the inverse of the (n x n) covariance matrix $\boldsymbol{\Lambda}_{\mathbf{WLS}}\left(\sigma^2_{\delta,B-WLS}\right)$ described in Equation 4.5, in which $\sigma^2_{\delta,B-WLS}$ is the mean model error variance estimated using B-WLS.

4.3.3 Influence

The influence measure, $\mathbf{D}^*$, for a GLS analysis as proposed by Tasker and Stedinger [1989, eq 25-26] is a generalized form of the Cook's D, and was computed as

$$D^*_i = \frac{1}{k}\frac{\left[\mathbf{H}^*\left(\boldsymbol{\Lambda}_{GLS}\left(\sigma^2_{\delta,MM-GLS}\right)\right)\right]_{ii}\hat{\varepsilon}^2_i}{\left[\left(\mathbf{I}-\mathbf{H}^*\right)\left(\boldsymbol{\Lambda}_{GLS}\left(\sigma^2_{\delta,MM-GLS}\right)\right)\right]^2_{ii}} \tag{4.16}$$

where k is the number of estimated regression coefficients, $\hat{\varepsilon}_i$ is the residual error for site i, $\mathbf{H}^*$ is an (n x n) matrix of the GLS leverage, $\boldsymbol{\Lambda}_{GLS}\left(\sigma^2_{\delta,MM-GLS}\right)$ is an (n x n) covariance matrix, $\mathbf{I}$ is an (n x n) identity matrix. Equation 4.16 can be simplified to

$$D^*_i = \frac{h^{'}_{ii}\hat{\varepsilon}^2_i}{k\left(\lambda^{'}_{ii}-h^{'}_{ii}\right)^2} \tag{4.17}$$

where $h^{'}_{ii}$ are the diagonal elements of

$$\mathbf{H}^{'} = \mathbf{H}^*\left(\boldsymbol{\Lambda}_{GLS}\left(\sigma^2_{\delta,MM-GLS}\right)\right) = \mathbf{X}\left\{\mathbf{X}^T\boldsymbol{\Lambda}^{-1}_{GLS}\left(\sigma^2_{\delta,MM-GLS}\right)\mathbf{X}\right\}^{-1}\mathbf{X}^T \tag{4.18}$$

and $\lambda_{ii}^{'}$ is the *i*th diagonal element of $\mathbf{\Lambda}_{GLS}\left(\sigma^2_{\delta,MM-GLS}\right)$.

The influence metric adopted by Tasker and Stedinger [1989] needs to be recast for the WLS/GLS methodology used in this study. Here the regression coefficients are estimated using WLS, whereas the precision of those coefficients and the precision of the model are calculated using B-GLS.

As shown in Equation 4.16, Cooks D contains two terms. The first describes the leverage of a point, which is measured as $Var\left[\hat{\gamma}_i \mid WLS \text{ model}\right]/Var\left[\hat{\varepsilon}_i \mid WLS \text{ model}\right]$, and the second is the square of the residual error divided by its variance.

The values of the required variance are provided below. In the following formulation $\mathbf{\Lambda} = \mathbf{\Lambda}_{GLS}\left(\sigma^2_{\delta,B-GLS}\right)$, $\mathbf{L} = \mathbf{\Lambda}_{WLS}\left(\sigma^2_{\delta,B-WLS}\right)$ and $\mathbf{H}^*_{WLS}$ = WLS/GLS Leverage (see Equation 4.15). This is done to simplify the following equations.

$$\begin{aligned} Var\left[\hat{\gamma} \mid WLS \text{ model}\right] &= \left(\mathbf{H}^*_{WLS}\right)\mathbf{L}\left(\mathbf{H}^{*T}_{WLS}\right) \\ &= \mathbf{X}\mathbf{W}_{WLS}\mathbf{L}\mathbf{W}^T_{WLS}\mathbf{X}^T \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{L}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{L}^{-1}\mathbf{L}\mathbf{L}^{-1}\mathbf{X}\left(\mathbf{X}^T\mathbf{L}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{L}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{X}\mathbf{W}\mathbf{L} = \left(\mathbf{H}^*_{WLS}\right)\mathbf{L} \end{aligned} \tag{4.19}$$

$$\begin{aligned} Var\left[\hat{\varepsilon} \mid WLS \text{ model}\right] &= \mathbf{E}\left\{\left(\boldsymbol{\gamma} - \mathbf{H}^*_{WLS}\hat{\boldsymbol{\gamma}}\right)\left(\boldsymbol{\gamma} - \mathbf{H}^*_{WLS}\hat{\boldsymbol{\gamma}}\right)^T\right\} \\ &= \mathbf{L} - \left(\mathbf{H}^*_{WLS}\right)\mathbf{L} - \mathbf{L}\left(\mathbf{H}^*_{WLS}\right)^T + \left(\mathbf{H}^*_{WLS}\right)\mathbf{L}\left(\mathbf{H}^*_{WLS}\right)^T \end{aligned} \tag{4.20}$$

$$\begin{aligned} Var\left[\hat{\varepsilon} \mid GLS \text{ model}\right] &= \mathbf{E}\left\{\left(\boldsymbol{\gamma} - \mathbf{H}^*_{WLS}\hat{\boldsymbol{\gamma}}\right)\left(\boldsymbol{\gamma} - \mathbf{H}^*_{WLS}\hat{\boldsymbol{\gamma}}\right)^T\right\} \\ &= \mathbf{\Lambda} - \left(\mathbf{H}^*_{WLS}\right)\mathbf{\Lambda} - \mathbf{\Lambda}\left(\mathbf{H}^*_{WLS}\right)^T + \left(\mathbf{H}^*_{WLS}\right)\mathbf{\Lambda}\left(\mathbf{H}^*_{WLS}\right)^T \end{aligned} \tag{4.21}$$

$$D_i^{WG} = \left(\frac{1}{k}\right)\left(\frac{Var\left[\hat{\gamma}_i \mid WLS \text{ model}\right]}{Var\left[\hat{\varepsilon}_i \mid WLS \text{ model}\right]}\right)\left(\frac{\varepsilon_i^2}{Var\left[\hat{\varepsilon}_i \mid GLS \text{ model}\right]}\right)$$
$$= \left(\frac{1}{k}\right)\left(\frac{h^*_{WLS,ii}}{1-h^*_{WLS,ii}}\right)\left(\frac{\varepsilon_i^2}{Var\left[\hat{\varepsilon}_i \mid GLS \text{ model}\right]}\right) \quad (4.22)$$

here $h^*_{WLS,ii}$ are the diagonal elements of H^*_{WLS}. The influence metric described in Equation 4.17 takes into account the mixed WLS/GLS analysis used to generate the regional skewness model. The predicted regional skewness model is estimated using WLS, and thus the leverage metric reflects the WLS weights that depend upon the diagonal covariance matrix. However, GLS describes the actual precision of the model and as well as the precision of the residuals. Thus, the last term in Equation 4.14 uses the correct estimate of the variance of the computed residuals, as computed by the GLS analysis.

If $\hat{\boldsymbol{\beta}}$ has dimensionality *k* and *N* is the sample size (number of basins in the study), then leverage values have a mean of *k/N*, and values greater than *2k/N* can be generally considered large. Influence values greater than 4/*N* are considered large [Tasker and Stedinger, 1989; Veilleux, 2009].

4.3.4 Pseudo R^2_δ

Pseudo R^2_δ describes the true fraction of the variability of the skewness coefficient from watershed-to-watershed explained by the fitted model. If $\sigma^2_\delta(k)$ is the B-GLS estimate of the model error variance for a model with *k* parameters (including the constant) and $\sigma^2_\delta(1)$ is the B-GLS estimate of the model error variance for the constant model, then Pseudo R^2_δ is computed as

$$R^2_\delta = 1 - \left[\sigma^2_\delta(k) / \sigma^2_\delta(1)\right] \quad (4.23)$$

4.4 Application of WLS/GLS Regression Framework to Develop a Regional Skewness Model for the Southeastern U.S.

4.4.1 Summary of the Southeastern U.S. Data and B-GLS Regional Skewness Model

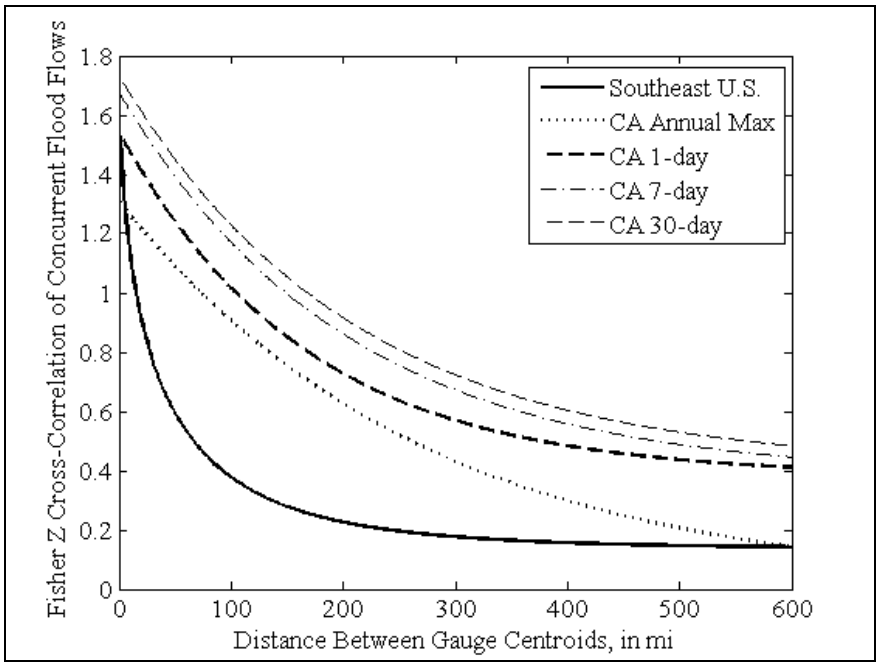
A regional skewness model for the Southeastern U.S. was developed recently using a Bayesian-GLS framework [Veilleux, 2009; Gotvald *et al*., 2009; Feaster *et al.*, 2009; Weaver *et al*., 2009]. The study was based upon annual peak flows from 342 stream flow gauges spread across seven states in the Southeastern United States. They were recommended by the United States Geological Surveys (USGS) Water Science Centers responsible for three states, Georgia, North Carolina, and South Carolina. In addition to the peak flow data, basin characteristics for the 342 sites were provided by the USGS Water Science Centers. The basin characteristics include percent of basin contained within physiographic provinces, as well as the more standard characteristics such as location of basin centroid, drainage area, main channel slope, and basin elevation. A cross-correlation model of the annual peak flows was developed which relates the Fisher Z transformation of the sample correlation to the distance between basin centroids. Based upon a Bayesian-GLS analysis of the 342 stations, a constant generalized regional skew model, $\hat{\gamma} = -0.019$, was selected for the Southeastern U.S. with an MSE = 0.14.

4.4.2 Comparison of Cross-Correlations of Flows in the Southeastern U.S. and California

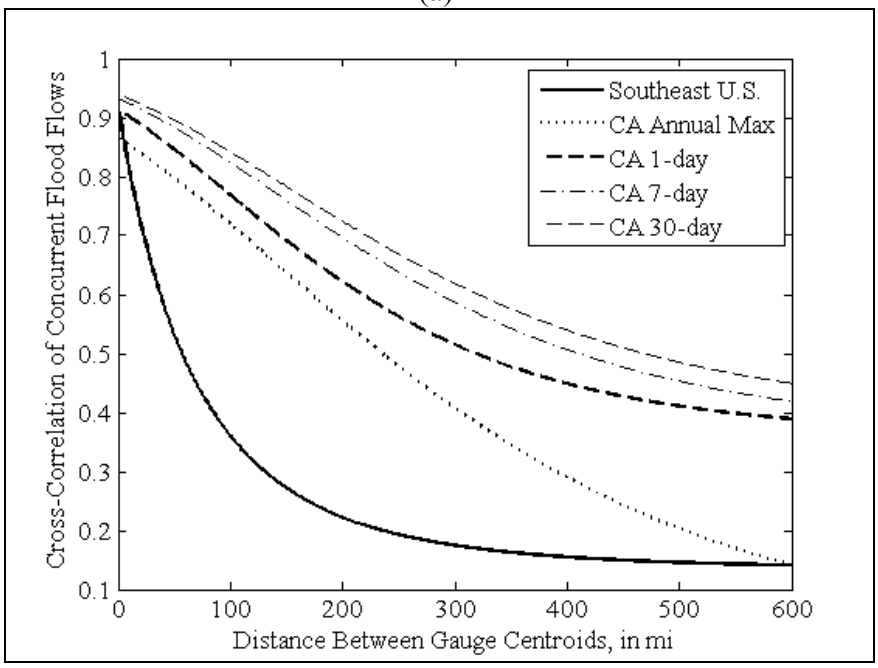
As documented in the Section 4.1.1, the new B-WLS/B-GLS regression framework was developed to deal with those regions with large cross-correlations among flood flows. As discussed in Section 4.1, large cross-correlations present difficulties in the regional skewness analysis because a B-GLS analysis seeks to exploit the cross-correlations among the sample skewness estimates to obtain the best possible estimates of the model parameters. If the cross-

correlations are large, the GLS estimators can become relatively complicated as a result of the effort to find the most efficient estimator of the parameters. Unfortunately, the precision of the cross-correlation estimates between any two particular sites is not of sufficient precision to justify the sophisticated weights (both positive and negative) that the B-GLS analysis generates.

Two regional skew studies in the State of California lead to the realization that the B-GLS method would need to be altered to properly deal with regions whose flood flows are highly correlated. The first study used annual maximum floods [Parrett *et al*., 2011] while the second study considered rainfall flood volumes corresponding to 1 to 30 day durations [Lamontagne *et al*., 2011]. These rainfall flood volumes exhibited even larger cross-correlations than did concurrent annual maxima. Figure 4.1 contains the cross-correlations between flood flows and distance between basin centroids for the Southeastern U.S. and the State of California.



(a)



(b)

Figure 4.1: Graphs of the cross-correlation of peak flows versus distance between basin centroids in miles. The solid black line is the cross-correlation function used in the B-GLS Southeastern U.S. regional skew study [Veilleux, 2009]. The dotted line is the cross-correlation function used the California annual maximum regional skew study [see Chapter 3 and Parrett *et al*., 2011]. The three dashed lines are from the California rainfall flood volume regional skew study [Lamontagne *et al.,* 2011]. Figure 4.1a graphs the Fisher Z transformed cross-correlations on the *y*-axis, while Figure 4.1b graphs the sample cross-correlations on the *y*-axis.

As shown in Figure 4.1, the cross-correlation functions of concurrent flood flows in California are very different from the one fitted to the Southeastern U.S. All of the cross-correlation functions show the same trend that the cross-correlation between concurrent flood flows decreases as the distance between gauge centroids increases. This is consistent with hydrological expectations; basins that are farther apart tend to experience different storm systems, climates, and geology resulting in decreased correlation between concurrent flood flows. However, the shape of the cross-correlation functions for flood flows in California yield much larger cross-correlations, especially at shorter distances. For example, when the distance between basin centroids is 30 miles the cross-correlations are 0.65 and 0.83 for the Southeastern U.S. and the California annual maximum study, respectively. When the distance between basin centroids is 150 miles, the difference between cross-correlations increases (the cross-correlations are 0.27 and 0.64 for the Southeastern U.S. and the California annual maximum study, respectively). These differences between the Southeastern U.S. and the California annual maximum study are large, and these differences only increase when the Southeastern U.S. function is compared to the California flood volume functions. Thus, Figure 4.1 illustrates the large difference between cross-correlation functions in the Southeastern U.S. and California, as well as, the large cross-correlations experienced in California.

4.4.3 Validating the WLS/GLS Methodology Using Annual Peak Flood Flow Data from the Southeastern U.S.

As discussed in the previous section, the Southeastern U.S. data set does not have the high cross-correlations that are present in California, and is believed to not encounter the same difficulties when using B-GLS to develop a regional skewness model. Thus by comparing

regional skewness models developed using B-GLS with those developed using the alternative WLS/GLS methodology presented in this chapter, the WLS/GLS methodology can be checked. This will help to confirm that the WLS/GLS performs similarly to the B-GLS methodology while avoiding the difficulties encountered by B-GLS in regions with large cross-correlations.

The results for the Southeastern U.S. regional skew models are presented in Table 4.1. OLS, Method-of-Moments WLS, Method-of-Moments GLS, and B-WLS regression analysis are included for comparison with both the B-GLS and B-WLS/B-GLS results, as shown in Table 4.1.

Based on the B-GLS regional skewness study results provided by Veilleux [2009], the use of explanatory variables did not result in a major improvement in the fit (Pseudo $R_{\delta}^{2} < 10\%$) as compared to the constant model, while adding to the complexity of the model. However, for the purposes of illustrating the methodology, the regression results are provided for both the Constant Model, as well as, Model H (which includes a regression constant and two physiographic region parameters: Blue Ridge and Sand Hills). The following sections describe the results in detail.

Table 4.1: Regional skew regression results for the Southeastern U.S. data set (number of sites = 342). Standard deviations are presented in parentheses ( ), $E\left(\sigma_\delta^2\right)$ is the expected value of model error variance, ASEV is the average sampling error variance, AVP_{new} is the average variance of prediction, and Effective Record Length (ERL) is a description of the information contained in the regional model estimates of the log-space skewness coefficient.

		Regression Parameters							Nominal
Method	**Model**	**Constant**	**Blue Ridge**	**Sand Hills**	$E\left(\sigma_\delta^2\right)$	**ASEV**	**AVPnew**	R_δ^2	**ERL**
OLS	Constant	-0.007 (0.035)	-	-	0.412 -	0.001	0.413	0%	25
	H	-0.104 (0.041)	0.290 (0.090) 0.2%	0.625 (0.170) 0.1%	0.391	0.003	0.394	5%	26
MM-WLS	Constant	0.001 (0.033)			0.223	0.001	0.224	0%	39
	H	-0.092 (0.040)	0.267 (0.086) 0.3%	0.605 (0.171) 0.1%	0.204	0.003	0.207	9%	41
MM-GLS	Constant	-0.017 (0.064)	-	-	0.178	0.004	0.182	0%	45
	H	-0.098 (0.067)	0.333 (0.119) 0.7%	0.525 (0.175) 0.4%	0.164	0.007	0.171	8%	48
B-WLS	Constant	0.001 (0.033)	-	-	0.208 (0.029)	0.001	0.209	0%	41
	H	-0.090 (0.040)	0.263 (0.088) 0.2%	0.597 (0.178) 0.0%	0.189 (0.034)	0.003	0.192	9%	44
B-GLS	Constant	-0.019 (0.063)	-	-	0.139 (0.021)	0.004	0.143	0%	55
	H	-0.099 (0.066)	0.335 (0.114) 0.3%	0.522 (0.167) 0.2%	0.129 (0.020)	0.007	0.136	7%	60
B-WLS/B-GLS	Constant	0.001 (0.068)			0.138 (0.021)	0.005	0.143	0%	55
	H	-0.091 (0.070)	0.266 (0.132) 4.4%	0.604 (0.180) 0.1%	0.128 (0.020)	0.008	0.137	7%	57

As shown in Table 4.1, the statistics for the constant models produced by B-GLS and B-WLS/B-GLS are very similar. They both have mean regression constants near zero (one standard deviation above and below the means contain zero), almost identical model error variances (σ_δ^2 =0.139 for B-GLS, σ_δ^2 =0.138 for B-WLS/B-GLS), and identical AVPnew (=0.143). The B-GLS and B-WLS/B-GLS results for Model H are similar and neither analysis reports a significant improvement over the constant model (R_δ^2 <10). These results verify that for a regional regression in which large cross-correlations among concurrent flood flows do not exist, B-GLS and B-WLS/B-GLS produce comparable results. This validates the use of the B-WLS/B-GLS methodology on those regions for which B-GLS fails to produce defensible results, *i.e.* those regions with large cross-correlations among concurrent flood flows. The B-WLS/B-GLS methodology was applied to the California rainfall-flood volumes regional skew and the results are presented in Lamontagne *et al*. [2011].

4.4.4 Pseudo ANOVA, Leverage and Influence for B-WLS/B-GLS for Southeastern U.S.

Table 4.2 contains the pseudo ANOVA results for the B-WLS/B-GLS Constant model and Model H. The pseudo ANOVA table clearly demonstrates that for the Southeastern U.S. data set, the sampling error is larger than the model error. These results are analogous with those for the corresponding B-GLS models [Veilleux, 2009].

Table 4.2: Pseudo ANOVA table for Southeastern U.S. regional skewness regression models (Constant and Model H) produced by B-WLS/B-GLS as presented in Table 4.1.

Source	Degrees-of-Freedom			Equations	Sum of squares	
		Constant	H		Constant	H
Model	k	0	2	$n\left[\sigma_\delta^2(0)-\sigma_\delta^2(k)\right]$	0.0	3.5
Model Error	n-k-1	341	339	$n\sigma_\delta^2(k)$	47	44
Sampling Error	n	342	342	$\sum_{i=1}^{n} Var(\hat{y}_i)$	59	59
Total	2n-1	683	683	$n\sigma_\delta^2(0)+\sum_{i=1}^{n} Var(\hat{y}_i)$	107	107
EVR	$EVR=\frac{\text{SS(sampling error)}}{\text{SS(model error)}}=\frac{\text{tr}[\boldsymbol{\Sigma}(\hat{\mathbf{y}})]}{n\sigma_\delta^2(k)}$				1.3	1.4
MBV	$MBV^*=\frac{Var\left[b_0^{WLS}\mid GLS\ analysis\right]}{Var\left[b_0^{WLS}\mid WLS\ analysis\right]}=\frac{w^T\Lambda w}{\sum_{i=1}^{n} w_i}$ where $w_i=\frac{1}{\Lambda_{ii}}$				5.4	5.6
$\mathbf{R_\delta^2}$	$R_\delta^2=1-\frac{\sigma_\delta^2(k)}{\sigma_\delta^2(0)}$				0.0%	7.4%

Figure 4.2 displays the leverage and values for the B-WLS/B-GLS constant model for the Southeastern U.S. The 27 sites included in the figure have high influence and thus have an unusual impact on the fitted regression relationship. The sites are ordered, starting from the left, by decreasing influence, as it identifies those sites that had a large impact on the analysis. No sites in the regression had high leverage, the differences in leverage values for the constant model reflect the variation in record lengths among sites. Site 243 has the highest influence value due to its large residual, the fourth largest residual in the study (*i.e.* the fourth largest unbiased at-site skew = 1.93), and its small drainage area (80 mi^2) (which is smaller than the three other sites with larger residuals).

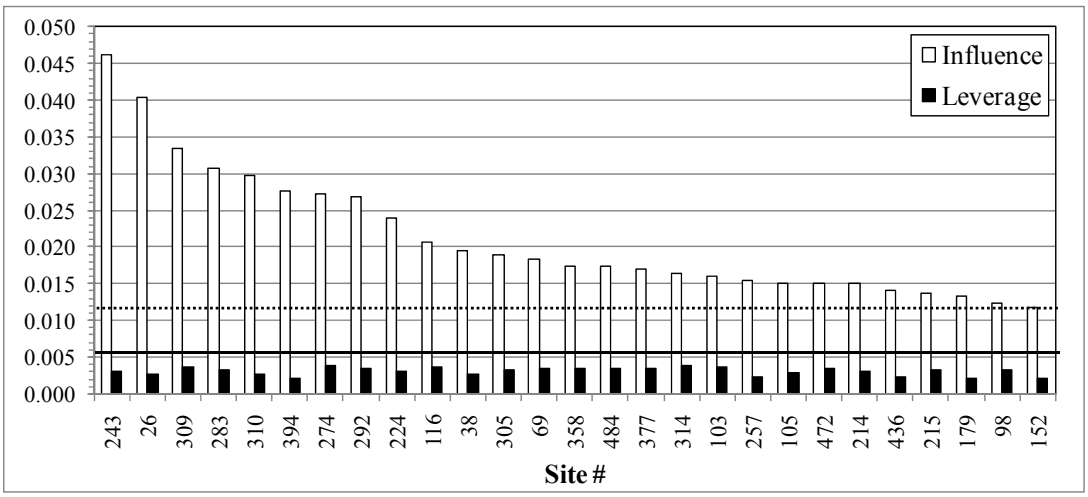


Figure 4.2: Regression Diagnostics: Leverage and influence for the Southeastern U.S. B-WLS/B-GLS Constant Model. The solid line represents the threshold for high leverage, while the dotted line represents the threshold for high leverage.

4.5 WLS/GLS Regional Skew Regression for the California Rainfall-Flood Volumes Data

This section provides a summary of the regional skew results for the California rainfall-flood volumes study developed by Lamontange [2011] and presented in Veilleux [2011]. The Bayesian WLS/GLS algorithm was applied to the California rainfall-flood data corresponding to different durations. The results for 1-, 3-, 7-, 15-, 30- days are presented in Table 4.3. The constant model had a relative small model error variance of prediction, and an expected variance of prediction ranging from 0.093 to 0.12. However, a nonlinear model based upon average basin elevation was much better with an average variance of prediction of 0.048 to 0.056 across the five durations. The non-linear relationship between average basin elevation and log-space skewness is shown in Figure 4.3 for the 3-day flow volume. The elevation effect reflects the impact of snow at the higher elevations, even though these flood volume series have been created to reflect only floods that were predominantly due to rain. At the higher elevations snowmelt can increase flood flows, while snow can also decrease the magnitude of flood peaks by capturing precipitation in the snowpack. Below 3,000 ft there is relatively little snow, and it is not

anticipated that elevation impacts regional skewness. Very few sites had basin average elevations above 5,000 feet, and those basins are clustered in the Southern Sierras representing the Kings, San Joaquin and Kern River watersheds. A linear model employing elevation was also considered; however it appeared to overestimate the impact of elevation on skewness for low and high elevation basins.

Table 4.3: Regional skewness models for California rainfall-flood volumes data. Standard deviations are presented in parentheses ( ), σ_δ^2 is the model error variance, Avg. SEV is the average sampling error variance, VP_{new} is the average variance of prediction, and Nominal ERL is a description of the information contained in the regional model estimates of the log-space skewness coefficient.

Duration	Type	B_0	B_1	$E[\sigma_\delta^2]$	Avg. SEV	VP_{new}	R_δ^2	Nominal ERL
1-Day	Constant	***-0.32***	-	0.078	0.035	0.113	0%	66
	NL Elev	-0.73	0.69	0.011	0.037	0.048	86%	150
		(0.22)	(0.12)	(0.01)				
3-Day	Constant	***-0.27***	-	0.080	0.039	0.118	0%	62
	NL Elev	-0.69	0.68	0.009	0.040	0.049	89%	143
		(0.22)	(0.11)	(0.01)				
7-Day	Constant	***-0.22***	-	0.053	0.040	0.093	0%	76
	NL Elev	-0.59	0.59	0.007	0.042	0.049	87%	140
		(0.23)	(0.11)	(0.01)				
15-Day	Constant	***-0.30***	-	0.034	0.043	0.076	0%	95
	NL Elev	-0.65	0.55	0.005	0.046	0.051	85%	141
		(0.24)	(0.11)	(0.00)				
30-Day	Constant	***-0.36***	-	0.033	0.044	0.076	0%	98
	NL Elev	-0.63	0.44	0.010	0.046	0.056	69%	133
		(0.24)	(0.11)	(0.01)				

*Non-significant terms in bold italics

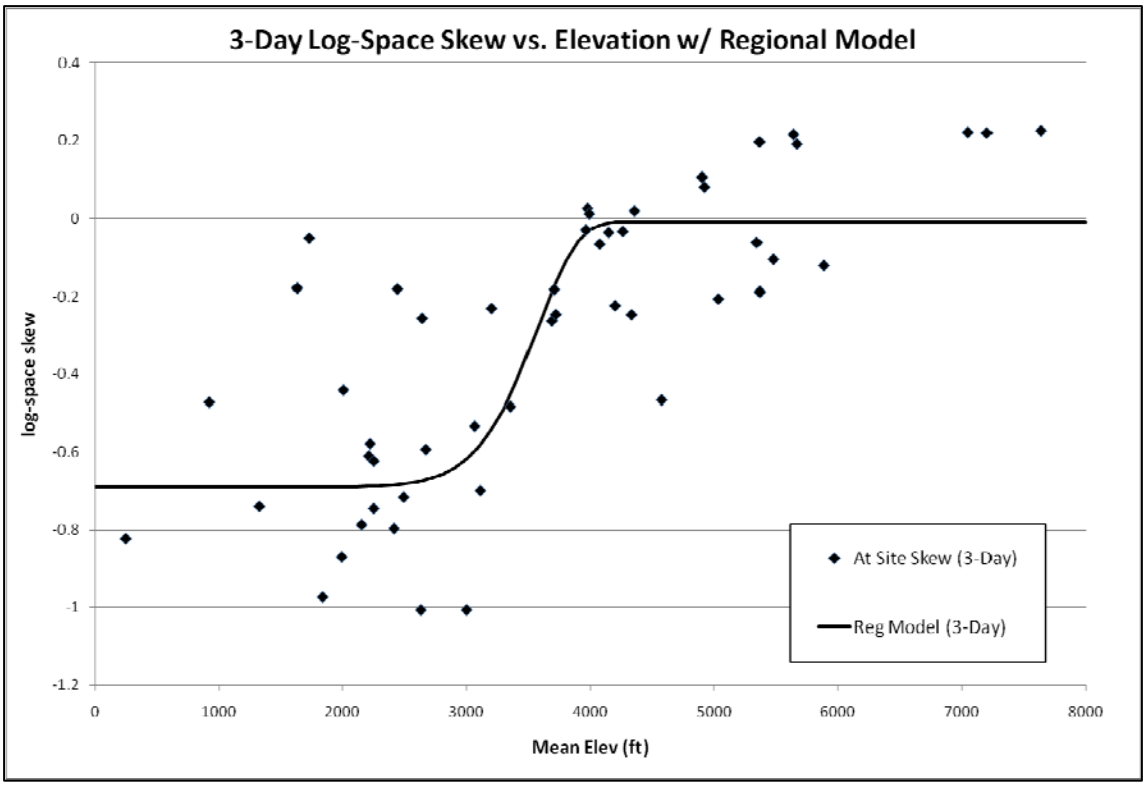


Figure 4.3: Log-space skewness versus average basin elevation (*ft*) for the 3-day flow volume. The black dots represent each of the 50 gage sites in the study, while the solid black line represents the non-linear elevation model from Table 1.

Leverage and Influence

Both leverage and influence values are computed for all sites for all durations. Leverage values did not change radically from one duration to another because the matrix of basin characteristics and the sample sizes were the same for all durations; however, the at-site skewness coefficients were different as were the model error variances, which resulted in some differences. On the other hand, the influence values depended upon the residuals computed from the individual skewness estimators for each duration, and thus changed from one duration to another. For this study, leverage values greater than 0.12 and influence values greater than 0.08 are considered to be large. None of the basins exhibits high leverage at any duration. Furthermore, no more than three basins exhibit high influence for any duration. Those basins

whose influence does exceed 0.08 do not exceed it by much, so their influences are not large enough to be alarming. Overall, this is an example wherein large leverage values were not expected. The nonlinear function of elevation ranged from a value of zero for basins below 3000 feet, to 1 for basins above 4200 feet. Thus it was impossible for any basin to have a particularly extreme value.

4.6 Conclusions

This chapter continues efforts to develop a spatial and regional statistical methodology for the estimation of hydrologic parameters. Regional log-space skewness studies (to support frequency analysis with the LP3 distribution) have furthered the development of the Bayesian-Generalized Least Squares methodology. Large cross-correlations between skewness coefficients in California require significant extensions of the B-GLS regression procedures published in Reis *et al*. [2005]. The new Bayesian WLS/GLS avoids instability problems that B-GLS encounters with such data sets. This chapter describes the B-WLS/B-GLS algorithm and equations for parameter estimation and diagnostic statistics, including Pseudo ANOVA, leverage, and influence. The Bayesian WLS/GLS methodology is used successfully to develop regional skewness models for the log-skew of the Southeastern U.S. annual maximum flows. The results validate that for this Southeastern U.S. annual maximum floods data set the B-WLS/B-GLS performs almost identically to that of the B-GLS analysis. The effective record length (ERL) of the regional skewness estimators for the Constant Model is about 55 years. These ERLs are better than the ERL of 17 years corresponding to the MSE reported for Plate 1 in Bulletin 17B, the current flood frequency guidelines for the United States.

REFERENCES

Feaster, T.D., Gotvald, A.J., and Weaver, J.C., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 3, South Carolina: U.S. Geological Survey Scientific Investigations Report, 2009-5156, 226 p.

Gotvald, A.J., Feaster, T.D., and Weaver, J.C., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 1, Georgia: U.S. Geological Survey Scientific Investigations Report, 2009–5043, 120 p.

Griffis, V.W., and Stedinger, J. R. , (2009), Log-Pearson type 3 distribution and its application in flood frequency analysis, III: sample skew and weighted skew estimators: Journal of Hydrology, v. 14, no. 2, p. 121–130.

Gruber, A.M., Reis, D.S., Jr., and Stedinger, J.R., (2007), Models of regional skew based on Bayesian GLS regression, Paper 40927-3285, *in* K.C. Kabbes ed., Restoring our Natural Habitat: Proceedings of the World Environmental and Water Resources Congress —, Tampa, Florida, American Society of Civil Engineers, May 15–18, 2007.

Interagency Advisory Committee on Water Data (1982), Guidelines for Determining Flood Flow Frequency, Bulletin #17B, U.S. Department of the Interior, U.S. Geological Survey, Office of Water Data Coordination, Reston Virginia.

Lamontagne, J., J. Stedinger, J. Ferris, D. Knifong, A. Veilleux, and D. Curry, (2011), Regional Skews for 1-Day, 3-Day, 7-Day, 15-Day, and 30-Day Duration Discharge for the Central Valley Region of California, Report Series XXXX-XXXX, U.S. Geological Survey (in press).

Martins, E.S., and J.R. Stedinger, (2002), Cross-correlation among estimators of shape, *Water Resources Research,* 38(11), doi: 10.1029/2002WR001589, 26 November.

Parrett, C., Veilleux, A., Stedinger, J.R., Barth, N.A., Knifong, D.L., and Ferris, J.C., (2011), Regional skew for California, and flood frequency for selected sites in the Sacramento–San Joaquin River Basin, based on data through water year 2006: U.S. Geological Survey Scientific Investigations Report 2010–5260, 94 p.

Reis, D.S., Jr., Stedinger, J.R., and Martins, E.S., (2005), Bayesian generalized least squares regression with application to the log Pearson type III regional skew estimation: Water Resources Research, 41, W10419, doi:10.1029/2004WR003445.

Tasker, G.D., and Stedinger, J.R., (1986), Regional skew with weighted LS regression: Journal of Water Resources Planning and Management, ASCE, v.112, no. 2, p. 225–237.

Tasker, G.D., and J.R. Stedinger, (1989), An Operational GLS Model for Hydrologic Regression, Journal of Hydrology, 111(1-4), 361–375.

Weaver, J.C., Feaster, T.D., and Gotvald, A.J., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 2, North Carolina: U.S. Geological Survey Scientific Investigations Report 2009–5158, 113 p.

Veilleux, A. G., (2009), Bayesian GLS Regression for Regionalization of Hydrologic Statistics, Floods and Bulletin 17 Skew, M.S. Thesis, Cornell University, August.

CHAPTER 5

CONCLUSIONS

5.1 Regional Hydrologic Regression Analysis

The research presented in this dissertation develops statistical techniques for estimating regional relationships for hydrologic parameters. These techniques include extensions of the Bayesian Generalizes Least Squares (B-GLS) framework presented in Reis *et al*. [2005]. Recent extensions include a Pseudo R_{δ}^{2}, pseudo Analysis of Variance (pseudo-ANOVA) table, plus a range of model performance, diagnostic and goodness-of-fit statistics. Particular attention is paid to development of leverage and influence metrics. In some cases, the B-GLS analysis proved to be unstable; for such cases, the research in this dissertation develops a stable Bayesian WLS/GLS procedure with the corresponding measures of precision, model performance, and diagnostic statistics.

Specifically, the research presented here extends the quasi-analytic Bayesian analysis of the Generalized Least Squares (GLS) regional hydrologic regression framework introduced by Reis *et al*. [2005], and furthered developed by Gruber *et al.* [2007], Gruber and Stedinger [2008], Parrett *et* al. [2011, appendices B and C], to estimate more accurately and precisely regional hydrologic relationships. In particular, examples in this dissertation consider estimation of a regional skewness coefficient relationship and its precision. Large cross-correlations among annual peak discharges, coupled with relatively small model error variances for skewness relationships, present difficulties for regional GLS skewness analyses. Problems arose because Bayesian GLS (B-GLS) estimators seek to exploit the cross-correlations among the sample skewness estimates to obtain the best possible estimators of the model parameters. However, if

the estimated cross-correlations are large, the B-GLS estimators can become relatively complicated as a result of the effort to find the most efficient estimator of the parameters. Unfortunately, it appears that the precision of the estimates of the cross-correlation between any two particular sites is not sufficiently precise to justify the complex weights (both positive and negative) that the B-GLS analysis generates. Thus, an alternate regression procedure using both Weighted Least Squares (WLS) and GLS is developed so that the regional skewness analysis can provide both stable and defensible results.

Special attention is devoted to the comparison of leverage and influence metrics for use with GLS regression. Leverage and influence metrics identify and should measure the potential and actual impact of unusual observations on fitted models. Derivations clarify the meaning of, and insight provided by, alternative measures of leverage and influence found in the literature and proposed herein. A related issue is the computation of the misrepresentation of beta variance (MBV) diagnostic statistic used to determine if a WLS regression is sufficient or if a GLS regression is needed. The formula for the MBV proposed by Griffis and Stedinger [2007] fails to do exactly what the authors intended. A revised MBV* statistic correctly computes the ratio of the precision of the constant term using a GLS analysis to that using a WLS analysis. In the examples considered, MBV and MBV* produce very similar values; while MBV provided the intended insight, it is recommend that future studies use MBV*.

5.2 United States Flood Flow Frequency Procedures and Regional Skew

Currently, *Bulletin 17B* allows for regional skew values to be obtained from the skewness coefficient map included with *Bulletin 17*, which was published in 1976. Because that map is over 35 years old, the regional skew values from the *Bulletin 17B* skew map do not reflect annual

maximum data acquired since 1976. Given concerns with climate change and hydrologic statistics, it makes little sense to make current decisions using a map that employed data that is at least 35 years old. The increase in available data since 1976, along with advances in computing power to support the Bayesian GLS regional hydrologic regression framework, allow for a much more precise estimate of a regional skewness coefficient and its accuracy for use in flood frequency analysis.

The recommended technique to perform flood frequency analyses, described in *Bulletin 17B*, is to fit a log-Pearson Type III (LP3) distribution to the series of annual maxima by the method of moments in log-space. The third moment of the LP3 distribution is the log-space skewness coefficient, which is very sensitive to extreme events, such as large or very small floods. Thus, an accurate estimate of the skewness coefficient is important in flood frequency analysis because the phenomena of the interest are large flood events which are represented by the upper tail of the distribution. Short record lengths at gauged sites make a regional estimate of skew extremely valuable in determining flood frequency estimates. This research focuses on advancing a procedure to develop regional skewness estimators for flood frequency analyses using a B-WLS/B-GLS regression framework. Special attention is also given to model performance and the meaning of, and insight provided by, alternative measures of leverage and influence.

Two examples of regionalization of the log-space skew illustrate the use of the B-WLS/B-GLS methodology. The first example, presented in Chapter 3, provides a regional skew analysis of annual peak flows from gauges in the State of California. The extended Bayesian-GLS methodology developed in this chapter provides stable and defensible results for the California regional skew analysis. The Bayesian-WLS analysis first developed estimators of the

regression coefficients for each regional skew model. By using WLS, the cross-correlations are not employed in estimating the regression coefficients. After the regression model coefficients were determined with WLS, the precision of a model and the precision of the regression coefficients are estimated using a modified GLS analysis. A Monte Carlo analysis determined the actual sample variance of the skewness coefficient when a low outlier test is employed to identify samples for special treatment. Finally, a modified Bayesian GLS analysis, using only data from pristine sites (*i.e*. sites without low outliers, zero flows, reconstructed records, or historical information), provided the estimate of the model error variance (the precision of the model) and the precision of the estimated parameters. This extended Bayesian-GLS methodology provided a stable regional skew model for California, while avoiding the instability issues encountered by the original B-GLS methodology due to the large cross-correlations between annual peak floods. The regional skew model recommended for the State of California has a Mean Square Error (MSE) equal to 0.14; the regional skew itself is dependent on the mean basin elevation

$$\hat{\gamma} = \beta_0 + \beta_2 \left[1 - \exp\left(-\left(Elev/6500 \right)^2 \right) \right] \tag{5.1}$$

This nonlinear elevation model with MSE = 0.14 is a definite improvement over the Bulletin 17B skew map, which reports MSE = 0.302. These results have been published by the USGS in Parrett *et al*. [2011].

A study of the hydrology of California confirms the results presented in Equation 5.1; the log-space skewness of annual maximum flood flow distributions in California are related to the mean basin elevations. Sites with mean basin elevations below 4,000 *ft* have their maximum annual floods driven by rainfall events. However, as the mean elevation of basins increases above 4,000 *ft,* the interaction of rainfall and snowmelt events increasingly effects the maximum

annual floods. As indicated by the regional skew model, this hydrology described through a nonlinear model of mean basin elevation, helps to explain variation in regional skewness values across California.

Chapter 3 provides a specialized Bayesian WLS/GLS methodology developed to estimate the parameters of a model for the skew of annual maximum flood series. Chapter 4 provides a more general and robust Bayesian WLS/GLS methodology to replace the standard B-GLS approach described in Chapter 2 when problems arise because of large cross-correlations. The first step in this improved Bayesian WLS/GLS methodology is to first estimate the regional log-space skew model using ordinary least squares (OLS). After the OLS regression parameters are estimated, they are then used to develop unbiased and relatively stable regional estimates of the skew for each gauge site. These regional at-site skew estimates are then used in conjunction with the at-site record lengths to estimate the sample variance of the unbiased skew estimator for each site. The second step is to perform a B-WLS analysis to obtain an estimate of the model error variance, and then subsequently to compute the WLS weights and to estimate the WLS regression parameters. The final step, is then to use the WLS weights in a B-GLS analysis to estimate the GLS model error variance as well as the precision of the regression coefficients obtained with the WLS weights. This improved Bayesian WLS/GLS methodology does not require a special Monte Carlo analysis as did the procedure developed in Chapter 3 for the data set available for that particular analysis, and thus, the computation of a regional regression model is much less complex. The improved Bayesian WLS/GLS procedure is used successfully to develop regional skewness models for the log-skew of the Southeastern U.S. annual maximum flows. The results validate that for this Southeastern U.S. annual maximum flood data set the B-WLS/B-GLS performs almost identically to that of the B-GLS analysis published in Veilleux

[2009], Feaster *et al*. [2009], Gotvald *et al*. [2009], and Weaver *et al*. [2009]. The effective record length (ERL) of the regional skewness estimators for the Constant Model is about 55 years. These ERLs are better than the ERL of 17 years corresponding to the MSE reported for Plate 1 in *Bulletin 17B*, the current flood frequency guidelines for the United States. The improved Bayesian WLS/GLS procedure was also used by Lamontagne *et al*. [2011] to develop regional models for floods of varying durations for the State of California.

In addition to developing the Bayesian WLS/GLS methodology for regional hydrologic regression, the research in this dissertation also provides a detailed review and development of diagnostic metrics for use with a GLS, a Bayesian GLS or a Bayesian WLS/GLS framework. Specifically, attention is focused on leverage and influence metrics. Through derivations and examples, the dissertation clarifies the information and insight provided by alternative leverage and influence metrics. Questions are raised regarding the strength of traditional leverage metrics which do not take into account the *x*-value at which the GLS regional regression model will be used to make a prediction. New x_0-leverage, which accounts for the characteristics (or *x*-values) at which a prediction will be made, is considered to be a more potentially informative metric. A new x_0–influence metric is also proposed. If the traditional leverage metric provides a poor representation of the leverage for an individual point, then the corresponding influence metric similarly will provide a poor representation of the impact of individual errors on the prediction made in different regions of the x-space. The use of leverage and influence in region-of-influence regression was also discussed.

The research documented in this dissertation shows that the improved Bayesian WLS/GLS regression model is an operational regional hydrologic regression methodology. In particular, examples are provided that illustrate the performance of the B-WLS/B-GLS analysis

in the estimation of regional skewness coefficients. In addition, the discussion and examples provided of leverage and influence metrics illustrates the information they provide and demonstrates their usefulness in identifying rogue observations and effectively addressing lack-of-fit.

5.3 Future Work

Much has been accomplished in the research reported in this dissertation. Future work should focus on further improvements in the methdology, and on employing the Bayesian GLS and Bayesian WLS/GLS methodology developed in this dissertation to estimate regional skew in other states, as well as to estimate regional skew on a national scale. Such studies would hopefully provide regional skewness estimators that would serve as a replacement to the 35-year old *Bulletin 17B* skew map and other previously conducted relatively naïve regional skew studies.

While the Bayesian WLS/GLS methodology has been shown to be operational in the studies discussed in this dissertation, it is expected that other complications will arise in performing regional skew regression for other parts of the United States. Work on a regional skew study for the State of Iowa has started and a new complication arose: many of the gauges in Iowa are crest stage gauges which only record flow values above an identified threshold. These crest stage gauges are different from continous gauges, which record almost all flow values, used in the previous regional skew studies. Thus, this complication will need to be addressed to generate a regional skew model for the State of Iowa.

Another interest would be a simple metric that recognized when the simpler B-GLS methodology is working well, and when the more stable B-WLS/B-GLS approach is needed.

Currently, other than comparing several analysis, or looking at the values of the weights assigned by WLS and GLS analyses, there is no simple criteria to indicate that a GLS analysis looks to be unstable.

Other areas of future methodological development could focus on developing improved crosss-correlation models of annual peak flows for use in the Bayesian WLS/GLS regional regression framework. The cross-correlation analyses presented in this dissertation as well as Tasker and Stedinger [1989], Reis *et al*. [2005], Veilleux [2009] and other studies use an OLS procedure. However, the OLS framework neglects the variation in the variances of the different residuals, and the cross-correlations among those estimators. Thus, the cross-correaltion analysis could be improved by implementing WLS and/or GLS analysis. In the studies cited, the cross-correlation analysis was only conducted on site pairs which had a substantial concurrent record length, so the validity of those particuar studies is not in question. But, there is an interest in being able to improve. It may also be possible to develop better predictor variables to include along with the distance between basin centroids, which has been employed in most recent studies.

Previous research reported in Veilleux [2009] attempted to identify redundant sites: pairs of sites corresponding to nested basins of roughly the same size. Additional work might consider both better screening metrics to identify potentailly redunant site pairs, as well as investigation of the important of identifing redundant sites and the consequence of including such pairs in an analysis. Not only do redundant sites introduce cross-correlation among the model errors, they can result in very large cross-correlations among concurrent annual flood peaks [Veilleux, 2009].

This disertation has provided both analytical analyses and examples exploring the information provided by both leverage and influence statistics for use with OLS, WLS and GLS

regression. Particular examples consider simple OLS data sets with a one- and two-dimensional x-space, as well as a simple GLS for a regression model with residuals arising from a time series model. Because of the wide-spread use of these statistics, clearly more work on these issues is warrented. Also of interest is the development of a conditional influence statistic and conditional residuals, where the magnitude of the residual errors is computed using its conditional mean given the observed value of other residuals.

This disertation represents tremendous strides in making the B-GLS procedure presented by Reis *et al*. [2005] and extended by Veilleux [2009] into an operational methodology, with several extensions that address situations where the simple B-GLS analysis becomes unstable. As discussed, other work remains to be done to further improved the approach and to provide better understand of the importance of different issues.

REFERENCES

Feaster, T.D., Gotvald, A.J., and Weaver, J.C., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 3, South Carolina: U.S. Geological Survey Scientific Investigations Report, 2009-5156, 226 p.

Gotvald, A.J., Feaster, T.D., and Weaver, J.C., (2009), Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 1, Georgia: U.S. Geological Survey Scientific Investigations Report, 2009–5043, 120 p.

Griffis, V. W., and J. R. Stedinger, (2007), The Use of GLS Regression in Regional Hydrologic Analyses, J. of Hydrology, 344(1-2), 82-95, [doi:10.1016/j.jhydrol.2007.06.023].

Gruber, Andrea M., Dirceu S. Reis Jr., and Jery R. Stedinger, [2007], Models of Regional Skew Based on Bayesian GLS Regression, Paper 40927-3285, World Environmental & Water Resources Conference - Restoring our Natural Habitat, K.C. Kabbes editor, Tampa, Florida, May 15-18.

Gruber, Andrea M. and Jery R. Stedinger, (2008), Models of LP3 Regional Skew, Data Selection and Bayesian GLS Regression, Paper 596, World Environmental and Water Resources Congress – Ahupua'a, Babcock, R.W. and R. Watson editors, Honolulu, Hawai'I, May 12-16.

Interagency Committee on Water Data (IACWD). (1982), *Guidelines for determining flood flow frequency: Bulletin 17B (revised and corrected)*, Hydrol. Subcomm., Washington, D.C., 28.

Lamontagne, J., J. Stedinger, J. Ferris, D. Knifong, A. Veilleux, and D. Curry, (2011), Regional Skews for 1-Day, 3-Day, 7-Day, 15-Day, and 30-Day Duration Discharge for the Central Valley Region of California, Report Series XXXX-XXXX, U.S. Geological Survey (in press).

Parrett, C., Veilleux, A., Stedinger, J.R., Barth, N.A., Knifong, D.L., and Ferris, J.C., (2011), Regional skew for California, and flood frequency for selected sites in the Sacramento–San Joaquin River Basin, based on data through water year 2006: U.S. Geological Survey Scientific Investigations Report 2010–5260, 94 p.

Reis, D. S., Jr., J. R. Stedinger, and E. S. Martins, (2005), Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation, *Water Resour. Res.*, *41*, W10419, doi:10.1029/2004WR003445.

Tasker, G.D., and J.R. Stedinger, (1989), An Operational GLS Model for Hydrologic Regression, Journal of Hydrology, 111(1-4), 361–375.

Veilleux, A. G. (2009), Bayesian GLS Regression for Regionalization of Hydrologic Statistics, Floods and Bullentin 17 Skew, M.S. Thesis, Cornell University, August.